Reinforcement Learning: Tutorial 10

Policy gradient methods: REINFORCE

Week 5
University of Amsterdam

Milena Kapralova
October 2024

Check-in

- How is it going?
- How is HW4?
- Are you ready to start HW5?
- If you have any feedback so far, please mail me at *m.kapralova@uva.nl*

Outline

1. Admin
2. Policy gradient methods: REINFORCE exercises
3. Ask anything about HW4 or 9.4 (HW5)
4. Try it yourself: Question 9.3

Admin

- Reminder that HW4 deadline is tomorrow @ 17:00
- Any questions?

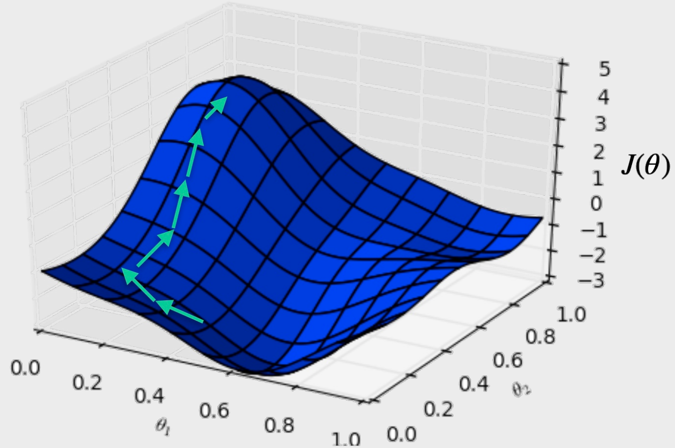
Tutorial 10 Overview

1. Policy gradient methods: REINFORCE exercises
2. Ask anything about HW4 or 9.4 (HW5)
3. Try it yourself: Question 9.3

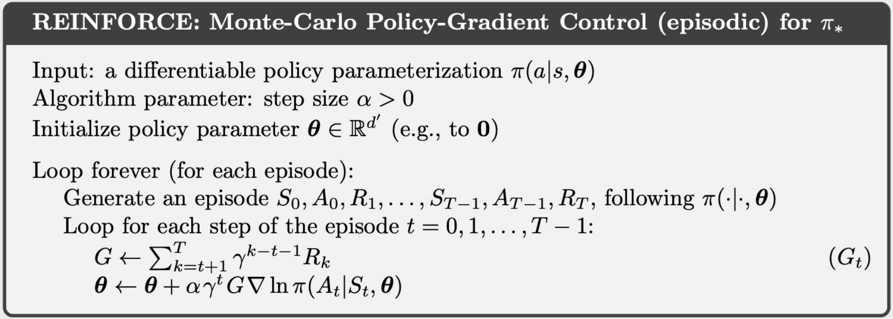
Tutorial 10 Overview

1. Policy gradient methods: REINFORCE exercises
   - Questions 9.1-9.2
2. Ask anything about HW4 or 9.4 (HW5)
3. Try it yourself: Question 9.3

Theory Intermezzo: How to find the best policy?



Theory Intermezzo: REINFORCE



REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for π_*

```
Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Algorithm parameter: step size $\alpha > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
    Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
    Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
        $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$                                        $(G_t)$
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t G \nabla \ln \pi(A_t|S_t, \boldsymbol{\theta})$
```

Q 9.1 REINFORCE v2 (Application)

We consider a simple example in a MDP setting with $\gamma = 1$, two states (s_0 and a terminal state T) and two actions $\{1, 2\}$. The agent followed the policy $\pi(a|s, \boldsymbol{\theta})$ and the episode $(s_0, a_0, -1, s_0, a_1, 2, T)$ was observed. We use the learning rate $\alpha = 1$ and initialize $\boldsymbol{\theta} = [1, 1]^\top$.

Q 9.1 REINFORCE v2 (Application)

We consider a simple example in a MDP setting with $\gamma = 1$, two states (s_0 and a terminal state T) and two actions $\{1, 2\}$. The agent followed the policy $\pi(a|s, \boldsymbol{\theta})$ and the episode $(s_0, a_0, -1, s_0, a_1, 2, T)$ was observed. We use the learning rate $\alpha = 1$ and initialize $\boldsymbol{\theta} = [1, 1]^\top$.

1. Assume that action probabilities are distributed following a categorical distribution with probabilities proportional to exponentiated weights:

$$\pi(a|s, \boldsymbol{\theta}) = \frac{e^{\phi(s,a)^\top \boldsymbol{\theta}}}{\sum_{b \in \{1,2\}} e^{\phi(s,b)^\top \boldsymbol{\theta}}},$$

$$\nabla_\theta \log \pi(a|s, \boldsymbol{\theta}) = \phi(s, a) - \sum_{b \in \{1,2\}} \pi(b|s, \boldsymbol{\theta}) \cdot \phi(s, b).$$

   And for state s_0 and actions, the feature vectors are $\phi(s_0, 1) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \phi(s_0, 2) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Please compute the update of $\boldsymbol{\theta}$ according to the REINFORCE (v2) update rule.

Q 9.1 REINFORCE v2 (Application)

We consider a simple example in a MDP setting with $\gamma = 1$, two states (s_0 and a terminal state T) and two actions $\{1, 2\}$. The agent followed the policy $\pi(a|s, \boldsymbol{\theta})$ and the episode $(s_0, a_0, -1, s_0, a_1, 2, T)$ was observed. We use the learning rate $\alpha = 1$ and initialize $\boldsymbol{\theta} = [1, 1]^\top$.

With the given episode, we have $G_0 = 1$ and $G_1 = 2$.

$$\sum_{b \in \{1,2\}} \pi(b|s_0, \boldsymbol{\theta}) \cdot \phi(s_0, b) = 0.5 \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 0.5 \cdot \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

$$\nabla_\theta \log \pi(a = 1|s, \boldsymbol{\theta}) = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$$

$$\nabla_\theta \log \pi(a = 2|s, \boldsymbol{\theta}) = \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix}$$

$$\boldsymbol{\theta} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha \cdot 1 \cdot \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix} + \alpha \cdot 2 \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}$$

Q 9.1 REINFORCE v2 (Application)

We consider a simple example in a MDP setting with $\gamma = 1$, two states (s_0 and a terminal state T) and two actions $\{1, 2\}$. The agent followed the policy $\pi(a|s, \boldsymbol{\theta})$ and the episode $(s_0, a_0, -1, s_0, a_1, 2, T)$ was observed. We use the learning rate $\alpha = 1$ and initialize $\boldsymbol{\theta} = [1, 1]^\top$.

2. Assume that the policy follows Gaussian distribution,

$$a \sim \mathcal{N}(\phi(s)^\top \boldsymbol{\theta}, 1),$$
$$\nabla_\theta \log \pi(a|s, \boldsymbol{\theta}) = (a - \phi(s)^\top \boldsymbol{\theta}) \cdot \phi(s).$$

Given $\phi(s_0) = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$, please compute the update of $\boldsymbol{\theta}$ according to the REINFORCE (v2) update rule.

Q 9.1 REINFORCE v2 (Application)

We consider a simple example in a MDP setting with $\gamma = 1$, two states (s_0 and a terminal state T) and two actions $\{1, 2\}$. The agent followed the policy $\pi(a|s, \boldsymbol{\theta})$ and the episode $(s_0, a_0, -1, s_0, a_1, 2, T)$ was observed. We use the learning rate $\alpha = 1$ and initialize $\boldsymbol{\theta} = [1, 1]^\top$.

$$\nabla_\theta \log \pi(a = 1|s, \boldsymbol{\theta}) = -0.5 \cdot \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} = \begin{bmatrix} -0.5 \\ -0.25 \end{bmatrix}$$

$$\nabla_\theta \log \pi(a = 2|s, \boldsymbol{\theta}) = \begin{bmatrix} 0.5 \\ 0.25 \end{bmatrix}$$

$$\boldsymbol{\theta} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha \cdot 1 \cdot \begin{bmatrix} -0.5 \\ -0.25 \end{bmatrix} + \alpha \cdot 2 \cdot \begin{bmatrix} 0.5 \\ 0.25 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 1.25 \end{bmatrix}$$

Q 9.1 REINFORCE v2 (Application)

We consider a simple example in a MDP setting with $\gamma = 1$, two states (s_0 and a terminal state T) and two actions $\{1, 2\}$. The agent followed the policy $\pi(a|s, \boldsymbol{\theta})$ and the episode $(s_0, a_0, -1, s_0, a_1, 2, T)$ was observed. We use the learning rate $\alpha = 1$ and initialize $\boldsymbol{\theta} = [1, 1]^\top$.

3. According to the updated $\boldsymbol{\theta}$ computed in above two questions, which action do you think will be more likely to be taken in the next episode starting from s_0? Why do you think that action should be taken?

Q 9.1 REINFORCE v2 (Application)

We consider a simple example in a MDP setting with $\gamma = 1$, two states (s_0 and a terminal state T) and two actions $\{1, 2\}$. The agent followed the policy $\pi(a|s, \boldsymbol{\theta})$ and the episode $(s_0, a_0, -1, s_0, a_1, 2, T)$ was observed. We use the learning rate $\alpha = 1$ and initialize $\boldsymbol{\theta} = [1, 1]^\top$.

3. According to the updated $\boldsymbol{\theta}$ computed in above two questions, which action do you think will be more likely to be taken in the next episode starting from s_0? Why do you think that action should be taken?

   In sub-question 1, $a = 2$ should be taken according to the updates. In sub-question 2, a bigger action should be taken.

   We observe positive reward when taking $a = 2$ and negative reward when taking $a = 1$. Intuitively, if the policy follows the category distribution, $a = 1$ could lead to larger reward. If the policy follows Gaussian distribution, we observe that the return for $a = 1$ is lower than the return for $a = 2$, and thus intuitively extrapolate that larger is better.

Q 9.2 REINFORCE v2 (Theory)

In this question, we will show that $\nabla_\theta J =$

$$\mathbb{E}_\tau\left[G(\tau)\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(\vec{a}_t|\vec{s}_t)\right] = \mathbb{E}_\tau\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(\vec{a}_t|\vec{s}_t)\sum_{t'=t+1:T} r_{t'}\right],$$

as claimed in the lecture (we focus on the non-discounted case). That is, that the gradient of the log probability of taking an action needs to be multiplied only with the rewards obtained after that action was executed.

Q 9.2 REINFORCE v2 (Theory)

In this question, we will show that $\nabla_\theta J =$

$$\mathbb{E}_\tau\left[G(\tau)\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(\vec{a}_t|\vec{s}_t)\right] = \mathbb{E}_\tau\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(\vec{a}_t|\vec{s}_t)\sum_{t'=t+1:T} r_{t'}\right],$$

as claimed in the lecture (we focus on the non-discounted case). That is, that the gradient of the log probability of taking an action needs to be multiplied only with the rewards obtained after that action was executed.

1. Show that:

$$\mathbb{E}_\tau\left[G(\tau)\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(\vec{a}_t|\vec{s}_t)\right] = \mathbb{E}_\tau\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(\vec{a}_t|\vec{s}_t)\sum_{t'=t+1:T} r_{t'}\right]$$
$$+ \underbrace{\sum_{t=1}^{T}\mathbb{E}_{\tau_{1:t-1},r_t}\mathbb{E}_{\vec{s}_t,\vec{a}_t|\tau_{1:t-1},r_t}\left[\nabla_\theta \log \pi_\theta(\vec{a}_t|\vec{s}_t)\sum_{t'=1:t} r_{t'}\right]}_{\text{term 2}}.$$

Q 9.2 REINFORCE v2 (Theory)

$$
\begin{aligned}
\nabla_\theta J &= \mathbb{E}_\tau \left[G(\tau) \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\vec{a}_t | \vec{s}_t) \right] \\
&= \mathbb{E}_\tau \left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\vec{a}_t | \vec{s}_t) \left(\sum_{t'=1:t} r_{t'} + \sum_{t'=t+1:T} r_{t'} \right) \right] \\
&= \mathbb{E}_\tau \left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\vec{a}_t | \vec{s}_t) \sum_{t'=t+1:T} r_{t'} \right] \\
&\quad + \sum_{t=1}^{T} \mathbb{E}_{\tau_{1:t-1}, r_t} \mathbb{E}_{\vec{s}_t, \vec{a}_t | \tau_{1:t-1}, r_t} \left[\nabla_\theta \log \pi_\theta(\vec{a}_t | \vec{s}_t) \sum_{t'=1:t} r_{t'} \right]
\end{aligned}
$$

In the first line we use the definition of G and split up the sum. To go to the second line, we distribute the sum of the two parts of the return and we use the identity in the hint to split up the expectation operator.

Q 9.2 REINFORCE v2 (Theory)

We split up the expectation as:

$$\mathbb{E}_{\tau_{1:T}}[\cdot] = \mathbb{E}_{\tau_{1:t-1},r_t}\mathbb{E}_{\vec{s}_t,\vec{a}_t|\tau_{1:t-1},r_t}\mathbb{E}_{\tau_{t+1:T}|\vec{s}_t,\vec{a}_t}[\cdot]$$

However, none of the terms depend on $\tau_{t+1:T}$, so we can drop the latter expectation of the three.

Q 9.2 REINFORCE v2 (Theory)

2. Show that the second term of equation in sub-question 1 is equal to 0. Hint: The main step is similar to how we showed the baseline does not introduce a bias in the lecture slides.

$$\nabla_\theta J = \mathbb{E}_\tau \left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\vec{a}_t | \vec{s}_t) \sum_{t'=t+1:T} r_{t'} \right]$$

$$+ \underbrace{\sum_{t=1}^{T} \mathbb{E}_{\tau_{1:t-1}, r_t} \left[\underbrace{\mathbb{E}_{\vec{s}_t, \vec{a}_t | \tau_{1:t-1}, r_t} \left[\nabla_\theta \log \pi_\theta(\vec{a}_t | \vec{s}_t) \right]}_{=0} \sum_{t'=1:t} r_{t'} \right]}_{=0}$$

Q 9.2 REINFORCE v2 (Theory)

We now pull in the expectation operator, since $r_{t'}$ doesn't depend on $\vec{s}_t, \vec{a}_t$ if $t' < t$. We have the same form as in the proof that using a baseline does not introduce bias: the expected value of a score function (gradient of a log probability), and we know this already to be 0. For completeness, this last fact can be written out as follows:

$$\begin{aligned}\mathbb{E}_{\vec{s}_t,\vec{a}_t|\tau_{1:t-1},r_t}\left[\nabla_{\vec{\theta}} \log \pi_{\vec{\theta}}(\tau)\right] &= \mathbb{E}_{\vec{s}_t,\vec{a}_t|\tau_{1:t-1},r_t}\left[\frac{\nabla_{\vec{\theta}}\pi(\tau)}{\pi(\tau)}\right] \\ &= \underbrace{\nabla_{\vec{\theta}} \underbrace{\mathbb{E}_{\vec{s}_t,\vec{a}_t|\tau_{1:t-1},r_t}[1]}_{=1}}_{=0} = 0\end{aligned}$$

WELL DONE!

Tutorial 10 Overview

1. Policy gradient methods: REINFORCE exercises
2. Ask anything about HW4 or 9.4 (HW5)
   - Questions 7.4, 8.3-8.4, 9.4
3. Try it yourself: Question 9.3

Ask anything about HW4 or 9.4 (HW5)

- 7.4: Theory
- 8.3: Theory
- 8.4: Coding (+ Little bit of theory)
- 9.4: Theory

Tutorial 10 Overview

1. Policy gradient methods: REINFORCE exercises
2. Ask anything about HW4 or 9.4 (HW5)
3. Try it yourself: Question 9.3

Q 9.3 Baseline and gradient variance

In the lecture, we have seen that introducing a constant baseline b for the trajectory reward $G(\tau)$ does not introduce a bias to our policy gradient.

$$\nabla J = \mathbb{E}_\tau\left[\Big(G(\tau) - b\Big)\nabla \log p(\tau)\right]$$

We now want to consider the variance when introducing a baseline.

Q 9.3 Baseline and gradient variance

In the lecture, we have seen that introducing a constant baseline b for the trajectory reward $G(\tau)$ does not introduce a bias to our policy gradient.

$$\nabla J = \mathbb{E}_\tau \left[\Big(G(\tau) - b \Big) \nabla \log p(\tau) \right]$$

We now want to consider the variance when introducing a baseline.

1. Derive the optimal constant baseline that minimizes the variance of the policy gradient. Interpret your result.
   Hint: First use the definition of variance to write out the variance of the gradient estimate. What should the derivative of this function w.r.t. b look like at optimality? Keep in mind the likelihood-ratio trick (Deisenroth et al, p.28).

Q 9.3 Baseline and gradient variance

$$
\begin{aligned}
\mathbb{V}\left[x\right] &= \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2 \\
\mathbb{V}\left[\nabla J\right] &= \mathbb{E}_\tau\left[\left(\left(G(\tau) - b\right)\nabla \log p(\tau)\right)^2\right] - \mathbb{E}_\tau\left[\left(G(\tau) - b\right)\nabla \log p(\tau)\right]^2 \\
\frac{d\mathbb{V}\left[\nabla J\right]}{\partial\ b} &= \frac{\partial}{\partial\ b}\left(\mathbb{E}_\tau\left[\left(G(\tau) - b\right)^2\left(\nabla \log p(\tau)\right)^2\right]\right. \\
&\quad \left. - \mathbb{E}_\tau\left[\left(G(\tau) - b\right)\nabla \log p(\tau)\right]^2\right) \\
&= \frac{\partial}{\partial\ b}\left(\mathbb{E}_\tau\left[\left(G(\tau) - b\right)^2\left(\nabla \log p(\tau)\right)^2\right]\right. \\
&\quad \left. - \left(\mathbb{E}_\tau\left[G(\tau)\nabla \log p(\tau)\right] - \mathbb{E}_\tau\left[b\nabla \log p(\tau)\right]\right)^2\right)
\end{aligned}
$$

Q 9.3 Baseline and gradient variance

$$\frac{d\mathbb{V}\left[\nabla J\right]}{\partial\ b} = \frac{\partial}{\partial\ b}\left(\mathbb{E}_\tau\left[\left(G(\tau) - b\right)^2\left(\nabla \log p(\tau)\right)^2\right]\right.$$

$$\left. - \left(\mathbb{E}_\tau\left[G(\tau)\nabla \log p(\tau)\right] - 0\right)^2\right)$$

$$= -2\ \mathbb{E}_\tau\left[\left(G(\tau) - b\right)\left(\nabla \log p(\tau)\right)^2\right]$$

Then the derivative w.r.t. b at optimality is:

$$\frac{d\mathbb{V}[\nabla J]}{\partial\ b} = 0 \iff b = \frac{\mathbb{E}_\tau\left[G(\tau)\left(\nabla \log p(\tau)\right)^2\right]}{\mathbb{E}_\tau\left[\left(\nabla \log p(\tau)\right)^2\right]}$$

The result says that the best baseline is the expected reward weighted by gradient magnitudes.

Q 9.3 Baseline and gradient variance

2. Consider the simple example in a bandit setting (i.e. no states):

$$\begin{aligned} r &= a + 2 \\ a &\sim \mathcal{N}(\theta, 1) \\ \nabla_\theta \log \pi(a) &= a - \theta \end{aligned}$$

Can you argue what should be the optimal constant baseline in this case?
Hint: Use your result from 1.

Q 9.3 Baseline and gradient variance

Using the result found in 1, plugging in the trajectory reward (note this example is stateless) and $\nabla_\theta \log\pi(a)$:

$$b = \frac{\mathbb{E}_\tau\left[G(\tau)\left(\nabla \log p(\tau)\right)^2\right]}{\mathbb{E}_\tau\left[\left(\nabla \log p(\tau)\right)^2\right]} = \frac{\mathbb{E}_a[(a+2)(a-\theta)^2]}{\mathbb{E}_a[(a-\theta)^2]}$$

$$b = \frac{\mathbb{E}_a[a^3 - 2a^2\theta + a\theta^2 + 2a^2 - 4a\theta + 2\theta^2]}{\mathbb{E}_a[a^2 - 2a\theta + \theta^2]}$$

$$b = \frac{\theta^3 + 3\theta - 2\theta(\theta^2+1) + \theta^3 + 2(\theta^2+1) - 4\theta^2 + 2\theta^2}{\theta^2 + 1 - 2\theta^2 + \theta^2}$$

$$b = \frac{\theta^3 + 3\theta - 2\theta^3 - 2\theta + \theta^3 + 2\theta^2 + 2 - 4\theta^2 + 2\theta^2}{1} = \theta + 2$$

In our example the best baseline is $b = \theta + 2$. The gradient weights cancel each other out due to the symmetry of the Gaussian policy and the quadratic term of the squared policy gradient centered at $a = \theta$.

Q 9.3 Baseline and gradient variance

3. Now consider a baseline that is not constant, but dependent on the state $b(s_t)$. We want to establish that in this case, the policy gradient remains unbiased. Show that

$$\mathbb{E}_\tau \left[\sum_{t=1}^{T} \nabla_\theta \log \pi(a_t|s_t) b(s_t) \right] = 0.$$

Hint: You can use the linearity of expectation or the law of iterated expectation to "decouple" the full trajectory τ in two parts, $s_1, a_1, \ldots, s_t$ and $a_t, r_{t+1}, s_{t+1}, \ldots$.

Q 9.3 Baseline and gradient variance

$$
\begin{aligned}
&\mathbb{E}_{\tau\sim\pi_\theta}\sum_{t=1}^{T}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)\right]\\
&=\sum_{t=1}^{T}\mathbb{E}_{s_{0:t},a_{0:t-1}}\left[\mathbb{E}_{s_{t+1:T},a_{t:T-1}}[\nabla_\theta \log \pi_\theta(a_t|s_t)b(s_t)\mid s_t]\right]\\
&=\sum_{t=1}^{T}\mathbb{E}_{s_{0:t},a_{0:t-1}}\left[b(s_t)\cdot\mathbb{E}_{s_{t+1:T},a_{t:T-1}}[\nabla_\theta \log \pi_\theta(a_t|s_t)\mid s_t]\right]\\
&=\sum_{t=1}^{T}\mathbb{E}_{s_{0:t},a_{0:t-1}}\left[b(s_t)\cdot\mathbb{E}_{a_t}[\nabla_\theta \log \pi_\theta(a_t|s_t)\mid s_t]\right]\\
&=\sum_{t=1}^{T}\mathbb{E}_{s_{0:t},a_{0:t-1}}\left[b(s_t)\cdot 0\right]=0
\end{aligned}
$$

Based on *this blogpost*.

That's it!

Good luck with the HW and see you on Monday