Reinforcement Learning: Tutorial 12

Advanced PS methods: NPG & TRPO

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. Advanced policy search exercises
3. Ask anything about HW5
4. Try it yourself: Question 11.3

Admin

- Please direct any questions about grading to Pieter Pierrot
- Reminder that HW5 deadline is tomorrow @ 17:00
- Any questions?

Tutorial 12 Overview

1. Advanced policy search exercises
2. Ask anything about HW5
3. Try it yourself: Question 11.3

Tutorial 12 Overview

1. Advanced policy search exercises
   - Questions 11.1-11.2
2. Ask anything about HW5
3. Try it yourself: Question 11.3

Theory Intermezzo: Updates

- The update for PG is $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\text{pg}} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
- The update for NPG is $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\text{npg}} \boldsymbol{F}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$
- The update for TRPO is $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\text{trpo}} \boldsymbol{F}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ where $\alpha_{\text{trpo}} = \sqrt{2 D_{KL} \left(\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})^T F \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \right)^{-1}}$

Q 11.1 Natural policy gradient

To explore the behavior of the gradients with different policy parametrizations we will use a stateless continuous bandit environment. In this environment, the agent performs a single action and receives a single reward before the episode is terminated. Furthermore, we assume that the reward function is known and the policy is represented by a normal distribution $\mathcal{N}(\mu(\theta_\mu), \sigma(\theta_\sigma))$.
In this exercise, we will consider a parametrization with hyperparameter k (we treat k as a design choice and thus its value cannot be changed during optimization):

$$r = a - a^2$$

$$\pi(a|\theta) = \frac{1}{\sigma(\theta_\sigma)\sqrt{2\pi}} \exp\left(- \frac{(a - \mu(\theta_\mu))^2}{2\sigma(\theta_\sigma)^2} \right)$$

$$\mu(\theta_\mu) = \theta_\mu$$

$$\sigma(\theta_\sigma) = k\theta_\sigma$$

Q 11.1 Natural policy gradient

$$r = a - a^2$$

$$\pi(a|\theta) = \frac{1}{\sigma(\theta_\sigma)\sqrt{2\pi}} \exp\left(- \frac{(a - \mu(\theta_\mu))^2}{2\sigma(\theta_\sigma)^2} \right)$$

$$\mu(\theta_\mu) = \theta_\mu$$

$$\sigma(\theta_\sigma) = k\theta_\sigma$$

Hint: For a Gaussian distribution $N(\mu, \sigma^2)$, we know the following expectations:

- $\mathbb{E}[a] = \mu = \theta_\mu$
- $\mathbb{E}[a^2] = \mu^2 + \sigma^2 = \theta_\mu^2 + \sigma(\theta_\sigma)^2$

1. For this problem we can calculate policy gradient analytically. Calculate the gradient of expected reward $\mathbb{E}_a[r]$ with respect to parameters θ_σ and θ_μ.

Q 11.1 Natural policy gradient

1. For this problem we can calculate policy gradient analytically. Calculate the gradient of expected reward $\mathbb{E}_a[r]$ with respect to parameters θ_σ and θ_μ.

$$\mathbb{E}_a[r] = \mathbb{E}_a\left[a - a^2\right] = \theta_\mu - {\theta_\mu}^2 - (k\theta_\sigma)^2$$
$$\nabla_{\theta_\mu}\mathbb{E}_a[r] = 1 - 2\theta_\mu$$
$$\nabla_{\theta_\sigma}\mathbb{E}_a[r] = -2k^2\theta_\sigma$$

Q 11.1 Natural policy gradient

2. In natural policy gradients, we take the update direction as $u = F^{-1}\nabla J(\theta)$. The fisher information matrix F and is then given by:

$$F_\theta = \mathbb{E}_a\Big[\nabla_\theta \log \pi(a|\theta)\nabla_\theta \log \pi(a|\theta)^T\Big]$$

The derivatives of the log probability wrt. parameters are:

$$\nabla_{\theta_\mu} \log \pi(a|\theta) = \frac{(a-\theta_\mu)}{(k\theta_\sigma)^2}$$

$$\nabla_{\theta_\sigma} \log \pi(a|\theta) = \frac{(a-\theta_\mu)^2}{k^2{\theta_\sigma}^3} - \frac{1}{\theta_\sigma}$$

Calculate the Fisher information matrix F_θ for our Gaussian policy.
Hint: Use results for central moments of a normal distribution:

$$\mathbb{E}_a\Big[(a-\theta_\mu)\Big] = 0, \quad \mathbb{E}_a\Big[(a-\theta_\mu)^2\Big] = \sigma(\theta_\sigma)^2$$

$$\mathbb{E}_a\Big[(a-\theta_\mu)^3\Big] = 0, \quad \mathbb{E}_a\Big[(a-\theta_\mu)^4\Big] = 3\sigma(\theta_\sigma)^4$$

Q 11.1 Natural policy gradient

Since we are in the stateless setting the expectation over the trajectory τ translates to a expectation over the action a. Additionally, since $\theta_\mu = \mu$ we can use identities for central moments of normal distribution.

$$F_{\theta_\mu,\theta_\mu} = \mathbb{E}_a\left[\left(\frac{(a-\theta_\mu)}{(k\theta_\sigma)^2}\right)^2\right] = \frac{\mathbb{E}_a\left[(a-\theta_\mu)^2\right]}{(k\theta_\sigma)^4} = \frac{(k\theta_\sigma)^2}{(k\theta_\sigma)^4} = \frac{1}{(k\theta_\sigma)^2}$$

$$F_{\theta_\sigma,\theta_\sigma} = \mathbb{E}_a\left[\left(\frac{(a-\theta_\mu)^2}{k^2{\theta_\sigma}^3} - \frac{1}{\theta_\sigma}\right)^2\right]$$

$$= \frac{\mathbb{E}_a\left[(a-\theta_\mu)^4\right]}{k^4{\theta_\sigma}^6} - 2\frac{\mathbb{E}_a\left[(a-\theta_\mu)^2\right]}{k^2{\theta_\sigma}^4} + \frac{1}{{\theta_\sigma}^2}$$

$$= \frac{3k^4{\theta_\sigma}^4}{k^4{\theta_\sigma}^6} - 2\frac{k^2{\theta_\sigma}^2}{k^2{\theta_\sigma}^4} + \frac{1}{{\theta_\sigma}^2} = \frac{3}{{\theta_\sigma}^2} - \frac{2}{{\theta_\sigma}^2} + \frac{1}{{\theta_\sigma}^2} = \frac{2}{{\theta_\sigma}^2}$$

Q 11.1 Natural policy gradient

$$
\begin{aligned}
F_{\theta_\mu,\theta_\sigma} = F_{\theta_\sigma,\theta_\mu} &= \mathbb{E}_a\left[\left(\frac{(a-\theta_\mu)}{(k\theta_\sigma)^2}\right)\left(\frac{(a-\theta_\mu)^2}{k^2{\theta_\sigma}^3} - \frac{1}{\theta_\sigma}\right)\right] \\
&= \frac{\mathbb{E}_a\left[(a-\theta_\mu)^3\right]}{k^4{\theta_\sigma}^5} - \frac{\mathbb{E}_a\left[(a-\theta_\mu)\right]}{k^2{\theta_\sigma}^3} \\
&= 0 - 0
\end{aligned}
$$

where we use the fact that the fourth central moment of the normal is $3\sigma^4$, the second central moment is σ^2 and third and first central moment are 0. Thus, the Fisher information matrix for our parameters $\theta^T = [\theta_\mu, \theta_\sigma]$ is

$$
F_\theta = \begin{bmatrix} \frac{1}{(k\theta_\sigma)^2} & 0 \\ 0 & \frac{2}{{\theta_\sigma}^2} \end{bmatrix}
$$

Q 11.1 Natural policy gradient

3. Consider two different parameterizations that represent the same policy $\mathcal{N}(0, 0.1)$:
   (a) $\theta_\mu = 0,\ \theta_\sigma = 1,\ k = 0.1$
   (b) $\theta_\mu = 0,\ \theta_\sigma = 0.01,\ k = 10$

   Perform a single gradient update with step size $\alpha = 1$ for both of them using natural policy gradient. Compare both policies. How is the result that you obtained related to the constraint of the natural policy gradient update?

Q 11.1 Natural policy gradient

Gradients from first sub-question:

$$\nabla_{\theta_\mu} \mathbb{E}_a[r] = 1 - 2\theta_\mu$$

$$\nabla_{\theta_\sigma} \mathbb{E}_a[r] = -2k^2\theta_\sigma, \qquad F_\theta^{-1} = \begin{bmatrix} (k\theta_\sigma)^2 & 0 \\ 0 & \frac{{\theta_\sigma}^2}{2} \end{bmatrix}$$

(a) $\theta_\mu = 0,\ \theta_\sigma = 1,\ k = 0.1$

$$\theta^* - \theta_0 \propto \begin{bmatrix} 0.01 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} 1 \\ -0.02 \end{bmatrix} = \begin{bmatrix} 0.01 \\ -0.01 \end{bmatrix}$$

$$\theta_\mu^a = 0.01 \qquad \theta_\sigma^a = 0.99$$

$$\mu^a = 0.01 \qquad \sigma^a = 0.099$$

Q 11.1 Natural policy gradient

Gradients from first sub-question:

$$\nabla_{\theta_\mu} \mathbb{E}_a[r] = 1 - 2\theta_\mu$$

$$\nabla_{\theta_\sigma} \mathbb{E}_a[r] = -2k^2\theta_\sigma, \qquad F_\theta^{-1} = \begin{bmatrix} (k\theta_\sigma)^2 & 0 \\ 0 & \frac{{\theta_\sigma}^2}{2} \end{bmatrix}$$

(b) $\theta_\mu = 0$, $\theta_\sigma = 0.01$, $k = 10$

$$\theta^* - \theta_0 \propto \begin{bmatrix} 0.01 & 0 \\ 0 & 0.00005 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 0.01 \\ -0.0001 \end{bmatrix}$$

$$\theta_\mu^b = 0.01 \qquad \theta_\sigma^b = 0.0099$$

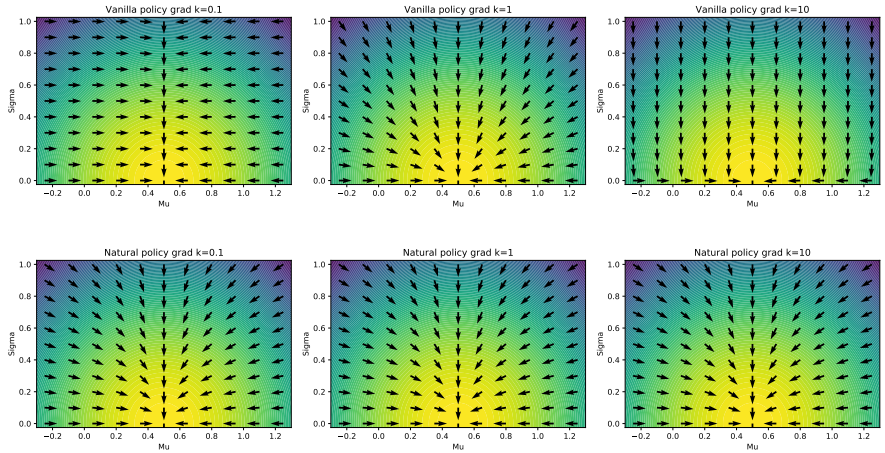
$$\mu^b = 0.01 \qquad \sigma^b = 0.099$$

Both policies are the same after the update. This is a direct consequence of the constraint used by the NPG which enforces the "difference" between current and new policy (KL divergence).

Q 11.1 Natural policy gradient

4. Plot the gradient directions with different $k = \{0.1, 1, 10\}$ for both policy gradient and natural policy gradient using the notebook *visualizing_gradients.ipynb* (you can find it on Canvas under Modules $\rightarrow$ Week 6). Would you expect both algorithms to work well if we use same step size α to update both σ and μ (for all k)? Would using separate step sizes α_μ and α_σ improve the performance of both algorithms?

Q 11.1 Natural policy gradient



For simple PG it will be difficult to use a single α for example for $k = 0.1$. In this case the gradient wrt. θ_μ will be much larger than the gradient wrt. θ_σ. High step size that would be suitable for θ_σ will cause θ_μ to oscillate and lower step size, suitable for θ_μ will lead to slow updates of θ_σ. In this case, having separate step sizes would be beneficial.

For NPG, a single step size would work well. A single step size is preferred because it is related to the KL-divergence constraint. Separate learning rates would likely not be beneficial.

Q 11.2 Trust region policy optimization

Policy gradient methods that use SGD to optimize a policy π_θ with parameters θ, depend implicitly on a linear approximation of the expected return $J(\theta)$. If we make gradient steps that are too large, this approximation fails, and we might not be able to find good policies.

Trust Region Policy Optimization (TRPO) is an algorithm that tries to make sure that we stay in a safe region around the current parameters, such that the approximation holds. In practice, it looks a lot like the natural gradient method. In this exercise, we assume a two-armed bandit setting with a Bernoulli policy, directly parameterised by θ:

$$\pi_\theta(a) = \theta^a(1-\theta)^{1-a}, \; a \in 0, 1$$
$$r(a) = a$$

Q 11.2 Trust region policy optimization

$$\pi_\theta(a) = \theta^a(1-\theta)^{1-a}, \; a \in 0, 1$$
$$r(a) = a$$

1. Let's start with the Natural Policy Gradient (NPG), which tries to compute a good update direction for the parameters. Compute this direction, given by $u = F^{-1}\nabla J(\theta)$, where F is the Fisher information matrix (here it's a 1D matrix), given by $F = -\mathbb{E}_a\left[\nabla^2_\theta \log \pi_\theta\right]$, and $J(\theta)$ is the expected return, given by $\mathbb{E}_a[r(a)]$.

Q 11.2 Trust region policy optimization

$$\pi_\theta(a) = \theta^a(1-\theta)^{1-a}, \; a \in 0, 1$$
$$r(a) = a$$

1. Let's start with the Natural Policy Gradient (NPG), which tries to compute a good update direction for the parameters. Compute this direction, given by $u = F^{-1}\nabla J(\theta)$, where F is the Fisher information matrix (here it's a 1D matrix), given by $F = -\mathbb{E}_a\left[\nabla^2_\theta \log \pi_\theta\right]$, and $J(\theta)$ is the expected return, given by $\mathbb{E}_a[r(a)]$.

   Note that $\pi_\theta(a=0) = 1-\theta$ and $\pi_\theta(a=1) = \theta$, so: $\nabla^2_\theta \log(1-\theta) = \nabla_\theta(\frac{-1}{1-\theta}) = \nabla_\theta(-1 \cdot (1-\theta)^{-1}) = (1-\theta)^{-2} \cdot \nabla_\theta(1-\theta) = \frac{-1}{(1-\theta)^2}$. Similarly, $\nabla^2_\theta \log(\theta) = \frac{-1}{\theta^2}$.

Q 11.2 Trust region policy optimization

$$
\begin{aligned}
F &= -\mathbb{E}_a \left[\nabla_\theta^2 \log \pi_\theta \right] \\
&= -\pi_\theta(a=0) \nabla_\theta^2 \log \pi_\theta(a=0) - \pi_\theta(a=1) \nabla_\theta^2 \log \pi_\theta(a=1) \\
&= (1-\theta) \frac{1}{(1-\theta)^2} + \theta \frac{1}{\theta^2} \\
&= \frac{1}{1-\theta} + \frac{1}{\theta} = \frac{1}{\theta(1-\theta)}
\end{aligned}
$$

$$
\begin{aligned}
\nabla J(\theta) &= \nabla \mathbb{E}_a \left[r(a) \right] \\
&= \nabla \left[(1-\theta) r(0) + \theta r(1) \right] = r(1) - r(0) = 1
\end{aligned}
$$

$$
u = F^{-1} \nabla J(\theta) = \theta(1-\theta)
$$

Q 11.2 Trust region policy optimization

2. Evaluate $u = F^{-1}\nabla J(\theta)$ for two settings of $\theta \in \{0.1, 0.5\}$. Then compute the KL divergence between the policy before and after the update suggested by the NPG (with a learning rate of 1). The KL divergence is given by $D_{KL}(\pi_{\theta_0}||\pi_\theta) = \mathbb{E}_{a\sim\pi_{\theta_0}}\left[\log \frac{\pi_{\theta_0}(a)}{\pi_\theta(a)}\right]$, where θ_0 are the old parameter values, and θ the updated values.

Q 11.2 Trust region policy optimization

2. Evaluate $u = F^{-1}\nabla J(\theta)$ for two settings of $\theta \in \{0.1, 0.5\}$. Then compute the KL divergence between the policy before and after the update suggested by the NPG (with a learning rate of 1). The KL divergence is given by $D_{KL}(\pi_{\theta_0}||\pi_\theta) = \mathbb{E}_{a\sim\pi_{\theta_0}}\left[\log \frac{\pi_{\theta_0}(a)}{\pi_\theta(a)}\right]$, where θ_0 are the old parameter values, and θ the updated values.

$\theta \in \{0.1, 0.5\}$ gives $u \in \{0.09, 0.25\}$. We thus have to compute the KL for $(\theta_0, \theta) = (0.1, 0.19)$ and $(\theta_0, \theta) = (0.5, 0.75)$. This evaluates to respective D_{KL}: $0.031, 0.144$, where we use the natural logarithm in the computation.

Q 11.2 Trust region policy optimization

3. The primary practical difference between TRPO and NPG, is that TRPO sets the maximum allowed step size of an update in addition to the direction (which is the same as the NPG direction). It does this by requiring that the norm of the NPG update is equal to some hyperparameter value, which itself may be interpreted as a desired D_{KL} between the policy before and after the parameter update. The TRPO update is $\theta = \theta_0 + \beta u$, where $\beta = \sqrt{2D_{KL}\,(u^T F u)^{-1}}$. For the initial value $\theta = 0.5$, what should the value of β be, if we want the update for $\theta = 0.5$ to have (approximately) the same D_{KL} as the update for $\theta = 0.1$ in step 2 of this exercise?

Q 11.2 Trust region policy optimization

3. The primary practical difference between TRPO and NPG, is that TRPO sets the maximum allowed step size of an update in addition to the direction (which is the same as the NPG direction). It does this by requiring that the norm of the NPG update is equal to some hyperparameter value, which itself may be interpreted as a desired D_{KL} between the policy before and after the parameter update. The TRPO update is $\theta = \theta_0 + \beta u$, where $\beta = \sqrt{2D_{KL}\,(u^T F u)^{-1}}$. For the initial value $\theta = 0.5$, what should the value of β be, if we want the update for $\theta = 0.5$ to have (approximately) the same D_{KL} as the update for $\theta = 0.1$ in step 2 of this exercise?

   From the previous question: the update for $\theta = 0.1$ has $D_{KL} = 0.031$. Furthermore, for $\theta = 0.5$, we can compute $u^T F u = \theta(1-\theta) = 0.25$. Then, in order to have this D_{KL} between the old and updated policy, the TRPO step-size equation requires that $\beta = \sqrt{2 \cdot 0.031/0.25} = \sqrt{0.248} \approx 0.498$.

Q 11.2 Trust region policy optimization

4. Compute D_{KL} for this update and compare to the update you found for $\theta = 0.1$ in question 2 of this exercise.

Q 11.2 Trust region policy optimization

4. Compute D_{KL} for this update and compare to the update you found for $\theta = 0.1$ in question 2 of this exercise.

   We update $\theta = 0.5 + \beta u = 0.5 + \sqrt{0.248} \cdot 0.25 \approx 0.63$. This gives a D_{KL} of approximately 0.035, which is very similar to what we found for $\theta = 0.1$ in question 2 of this exercise. Thus it indeed seems that the step size β - computed from the TRPO constraint - satisfies that our updates lead to changes in policy that are constrained by the chosen value for the KL divergence, independent of the parameter value we start at.


WELL DONE!

Tutorial 12 Overview

1. Policy gradient methods exercises
2. Ask anything about HW5
   - Questions 9.4, 10.3-10.4
3. Try it yourself: Question 11.3

Ask anything about HW5

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

Tutorial 12 Overview

1. Policy gradient methods exercises
2. Ask anything about HW5
3. Try it yourself: Question 11.3

Q 11.3 Update directions

Consider a game of rock, paper, scissors. The policy is parametrized by a Categorical distribution with parameters $\boldsymbol{\theta} = (\theta_{\text{rock}}, \theta_{\text{paper}})$ where each parameter corresponds to the probability of selecting the corresponding hand sign:

$$\begin{aligned} p(a = \text{rock}|\boldsymbol{\theta}) &= \theta_{\text{rock}} \\ p(a = \text{paper}|\boldsymbol{\theta}) &= \theta_{\text{paper}} \\ p(a = \text{scissors}|\boldsymbol{\theta}) &= 1 - \theta_{\text{paper}} - \theta_{\text{rock}}. \end{aligned}$$

Furthermore, the Fisher information matrix for the categorical distribution is given as:

$$\boldsymbol{F} = \begin{bmatrix} \frac{1}{\theta_{\text{rock}}} + \frac{1}{1-\theta_{\text{rock}}-\theta_{\text{paper}}} & \frac{1}{1-\theta_{\text{rock}}-\theta_{\text{paper}}} \\ \frac{1}{1-\theta_{\text{rock}}-\theta_{\text{paper}}} & \frac{1}{\theta_{\text{paper}}} + \frac{1}{1-\theta_{\text{rock}}-\theta_{\text{paper}}} \end{bmatrix}$$

Q 11.3 Update directions

Alice starts with a *uniform policy*. She samples several episodes, where each episode is a single game consisting of sampling a single action from the policy (i.e. rock, paper, scissors) and obtaining a reward of 1 (win),0 (draw) or -1 (loss) depending on the opponents action. The opponent can be seen as part of the environment. Alice performs an update based on the sampled games by using Natural Policy Gradient (NPG) and averaging the gradients over all episodes. She notices that after the update $\theta_{\text{rock}} > \theta_{\text{paper}}$.

You can assume that learning rates and target KL are > 0. *As part of your answer, give the update equation for each learning algorithm and use them to reason about your answer.*

Q 11.3 Update directions

1. Alice claims that NPG and TRPO would always obtain the same ordering regardless of the sampled action and outcome in the game (e.g. $\theta_{\text{rock}} > \theta_{\text{paper}}$ for both methods). Bob claims that the ordering could be different depending on the sampled game, learning rate of NPG and target KL used for TRPO. Who is right? Explain your answer.

Q 11.3 Update directions

1. Alice claims that NPG and TRPO would always obtain the same ordering regardless of the sampled action and outcome in the game (e.g. $\theta_{\text{rock}} > \theta_{\text{paper}}$ for both methods). Bob claims that the ordering could be different depending on the sampled game, learning rate of NPG and target KL used for TRPO. Who is right? Explain your answer.

   The update for NPG is $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\text{npg}} \boldsymbol{F}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$.
   The update for TRPO is $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\text{trpo}} \boldsymbol{F}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ where $\alpha_{\text{trpo}} = \sqrt{2 D_{KL} \left(\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})^T F \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \right)^{-1}}$.
   Since we have that $\theta_{\text{rock}} > \theta_{\text{paper}}$ after NPG update, it means that the same ordering applies for corresponding entries in $\boldsymbol{F}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$. Since the same $\boldsymbol{F}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is used in TRPO update, the only difference is the scaling by learning rate. The post-update ordering should therefore not change, no matter values we use because. Alice is right.

Q 11.3 Update directions

2. Alice claims that NPG and 'vanilla' policy gradients (PG) would always obtain the same ordering regardless of the sampled action and outcome in the game (e.g. $\theta_{\text{rock}} > \theta_{\text{paper}}$ for both methods). Bob claims that the ordering could be different depending on the sampled game and learning rates. Who is right? Explain your answer.

Q 11.3 Update directions

2. Alice claims that NPG and 'vanilla' policy gradients (PG) would always obtain the same ordering regardless of the sampled action and outcome in the game (e.g. $\theta_{\text{rock}} > \theta_{\text{paper}}$ for both methods). Bob claims that the ordering could be different depending on the sampled game and learning rates. Who is right? Explain your answer.

The update for NPG is $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\text{npg}} \boldsymbol{F}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$.
The update for PG is $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_{\text{pg}} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$.
Because we are using a uniform policy we have $\boldsymbol{F}^{-1} = \begin{bmatrix} \frac{2}{9} & -\frac{1}{9} \\ -\frac{1}{9} & \frac{2}{9} \end{bmatrix}$.The inverted Fisher information matrix is not a scaled identity matrix and will therefore change the direction of the update. Alice is wrong. Depending on the sampled games the ordering NPG and PG end up with can be different. Bob is right: The learning rate has no effect on the ordering.

Q 11.3 Update directions

3. In general, which updates (from PG, NPG, TRPO) update parameters in the same direction in the parameter space?

Q 11.3 Update directions

3. In general, which updates (from PG, NPG, TRPO) update parameters in the same direction in the parameter space?

   NPG and TRPO differ only by a scaling factor since they both use $\boldsymbol{F}^{-1}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$. They thus update the parameters in the same direction. However, PG does not use $\boldsymbol{F}^{-1}$ so the direction of the update is usually different from NPG and TRPO.

That's it!

Good luck with the HW and see you on Monday