UNIVERSITEIT VAN AMSTERDAM
Faculteit Natuurwetenschappen, Wiskunde en Informatica

Reinforcement Learning Week 5

By

Sergio Alejandro Gutierrez Maury, 11821353
Ilse Feenstra, 13628356

October 13, 2023

Homework: Limits of policy gradients

In this section we parameterize our policy with a univariate Gaussian probability density

$$N(\mu(\theta_\mu), \sigma(\theta_\sigma))$$

over real-valued actions. We consider a scenario with a single state where you can assume that an episode solely consists of one action and one reward. Mean and variance are learned:

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

(1)

1. (2 pts.) Calculate $\nabla \log \pi(a|\theta)$ w.r.t. parameters θ_σ and θ_μ for two different parametrizations. (When using normal distribution as a policy, it is common to parametrize standard deviation as an exponential, other parametrization is used for illustration purposes):

   (a) $\mu(\theta_\mu) = \theta_\mu$, $\sigma(\theta_\sigma) = \exp(\theta_\sigma)$

   (b) $\mu(\theta_\mu) = \theta_\mu$, $\sigma(\theta_\sigma) = \theta_\sigma^2$

First take the log of $\pi(a|s,\theta)$:

$$log\pi(a,\theta) = -log(\sigma(\theta_\sigma) - \frac{1}{2}log(2\pi) - \frac{(a-\theta_\mu)^2}{2\sigma(\theta_\sigma)^2}$$

a) Take the derivatives:

With respect to θ_μ:

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

With respect to θ_σ:

$$\frac{\partial \log \pi(a|\theta)}{\partial \theta_\sigma} = \frac{\partial \log \pi(a|\theta)}{\partial \sigma} * \frac{\partial \sigma}{\partial \theta_\sigma}$$

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

$$\frac{\partial \sigma}{\partial \theta_\sigma} = exp(\theta_\sigma)$$

$$\frac{\partial \log \pi(a|\theta)}{\partial \theta_\sigma} = \left(\frac{-1}{exp(\theta_\sigma)} + \frac{(a-\mu(\theta_\mu))^2}{\exp(\theta_\sigma)^3}\right) \times \exp(\theta_\sigma)$$

$$= -1 + \frac{(a-\mu(\theta_\mu))^2}{\exp(\theta_\sigma)^2}$$

b) Take the derivatives:

With respect to θ_μ:

$$\frac{\partial \log \pi(a|\theta)}{\partial \theta_\mu} = \frac{a-\theta_\mu}{\theta_\sigma^4} =$$

With respect to θ:

$$\frac{\partial \log \pi(a|\theta)}{\partial \theta_\sigma} = \frac{\partial \log \pi(a|\theta)}{\partial \sigma} * \frac{\partial \sigma}{\partial \theta_\sigma}$$

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

$$\frac{\partial \sigma}{\partial \theta_\sigma} = 2\theta_\sigma$$

$$\frac{\partial \log \pi(a|\theta)}{\partial \theta_\sigma} = \left(-\frac{1}{\theta_\sigma^2} + \frac{(a-\mu(\theta_\mu))^2}{\theta_\sigma^6}\right) \times 2\theta_\sigma$$

$$= -\frac{2}{\theta_\sigma} + \frac{2(a-\mu(\theta_\mu))^2}{\theta_\sigma^5}$$

2. (2 pts.) Suppose our current parameters are $\mu = 0$ and $\sigma = 4$. We now observe an episode with $a = 3$ and reward $r = 3$. Perform a gradient update on θ_μ and θ_σ with learning rate $\alpha = 0.1$ using the policy gradient for both parametrizations. What are the new parameter values? What are the new policies $N(\mu, \sigma)$?

For $\mu(\theta_\mu) = \theta_\mu$, $\sigma(\theta_\sigma) = \exp(\theta_\sigma)$:

$$\theta_\sigma = ln(4)$$

$$\theta_{t+1}^{\mu} = ln(4) + 0.1 * 3 * \nabla \log \pi(3|0)$$
$$= 0 + 0.3 * \frac{(3-0)}{exp(2ln(4))}$$

$$= \frac{0.9}{16}$$

$$\begin{aligned}\theta^{\sigma}_{t+1} &= ln(4) + 0.1 * 3 * \nabla \log \pi(3|ln(4)) \\ &= ln(4) + 0.3 * (-1 + \frac{(3-0)^2}{exp(2ln(4))}) \\ &= ln(4) + 0.3 * (-1 + \frac{9}{exp(8)}) = ln(4) - 0.3 + \frac{0.27}{16}\end{aligned}$$

For $\mu(\theta_\mu) = \theta_\mu$, $\sigma(\theta_\sigma) = \theta_\sigma^2$:

$$\theta_\sigma = 2$$

$$\begin{aligned}\theta^{\mu}_{t+1} &= 0 + 0.1 * 3 * \nabla \log \pi(3|0) \\ &= 0.3 * \frac{3-0}{2^4} = 0.3 * \frac{3}{16} = \frac{0.9}{16}\end{aligned}$$

$$\begin{aligned}\theta^{\sigma}_{t+1} &= 2 + 0.1 * 3 * \nabla \log \pi(3|2) \\ &= 2 + 0.3 * \left(-\frac{2}{2} + \frac{2(3-0)^2}{2^5}\right) \\ &= 2 + 0.3 * \left(-1 + \frac{18}{32}\right) = 1.7 + \frac{5.4}{32}\end{aligned}$$

3. (1 pts.) Use the results you obtained in the previous sub-question (updated policies) to explain a drawback of a simple policy gradient.

For both parameterizations, the θ_σ got larger. In this example, the θ_σ is the variance. While the reward was positive, the θ still increased, meaning a higher variance. This is a drawback of a simple policy gradient because this causes slow learning.

Homework: Coding Assignment - Policy Gradients

1. (1 pt.) We have spent a lot of time working on value based methods. We will now switch to policy based methods, i.e. learn a policy directly rather than learn a value function from which the policy follows. Mention two advantages of using a policy

based method.

The first advantage is that policy-based methods optimize the policy function directly without needing an intermediate value function. This can be particularly useful in certain problems where the policy might be simpler to represent or learn than the value function. For instance, in continuous action spaces, it might be easier to directly parametrize and adjust a policy rather than estimating a value for every possible action.

The second advantage is that policy-based methods can use stochastic policies. Unlike value-based methods, which usually result in deterministic policies, policy-based methods can naturally represent and optimize stochastic policies. This can be useful because a stochastic policy can naturally incorporate exploration without the need for an explicit exploration strategy like -greedy, which is often used with value-based methods.

2. Download the notebook `RLLab5 PG.zip` from canvas assignments and follow the instructions.

Done in Jupyter Notebook.

Homework: Update Directions

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

$$p(a = \text{rock}|\theta) = \theta_{\text{rock}} \tag{1}$$

$$p(a = \text{paper}|\theta) = \theta_{\text{paper}} \tag{2}$$

$$p(a = \text{scissors}|\theta) = 1 - \theta_{\text{paper}} - \theta_{\text{rock}}. \tag{3}$$

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

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

Alice starts with a uniform policy. She samples several games, where a single game consists 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.

1. (1.5 pts.) Alice claims that she would always obtain the same ordering $(\theta_{\text{rock}} > \theta_{\text{paper}})$ if she had used TRPO for the update instead (using the same data). Bob claims that the ordering could be different depending on the sampled games (different games but used for both NPG and TRPO), learning rate of NPG and target KL used for TRPO. Who is right? Explain your answer.

Alice is right because TRPO and NPG move in the same direction. If the gradient moves in the same direction, then the ordering will be the same for both methods.

2. (1.5 pts.) Alice also claims that she would always obtain the same ordering $(\theta_{\text{rock}} > \theta_{\text{paper}})$ if she had used Policy Gradient (PG) for the update instead (using the same data). Bob claims that the ordering could be different depending on the sampled data (different sample but used for both NPG and PG) and learning rates used for PG and NPG. Who is right? Explain your answer.

Bob is correct because PG and NPG do not have the same update. NPG use the $F^{-}1$ matrix to scale the gradient, while PG uses the raw gradient. F and the inverse of F are not the same. F is given. The inverse of F is:

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

Since this is not equal to the F given, F^{-1} and F are not the same. The fact that the inverse of F is not the same as F implies that the relative adjustment of gradients for θ_{rock} and θ_{paper} by NPG could be different from their raw values. Therefore, NPG and PG do not necessarily make the same update and do not have the same ordering.

3. (1 pt.) In general, which updates (from PG, NPG, TRPO) update parameters in the same direction in the parameter space?

NGP and TRPO move in the same direction because the direction depends on the Fisher diagonal, which is the same in the case of the TRPO and NGP. PG does not always have the same direction.