Reinforcement Learning: Homework 4

Jorrim Prins
Student nr.: 11038934
jorrimprins.prins@student.uva.nl

Yke Rusticus
Student nr.: 11306386
yke.rusticus@student.uva.nl

7.1 Coding Assignment - Policy Gradients

1. Policy based methods have multiple advantages over value based methods:

- Policy based methods are able to handle continuous actions by choosing to use a policy with continuous output (such as a linear function or a neural network).
- Policy based methods can ensure smoothness of the policies, as a small step size will automatically lead to a small policy change.
- Policy based methods are able to learn stochastic policies, while value based methods are not.

8.1 Natural policy gradient

1.

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

2. Fisher information matrix is given by:

$$F_\theta = \mathbb{E}_\tau \left[\nabla_{d\theta} \log \pi \left(a \mid \theta_0 + d\theta\right) \nabla_{d\theta} \log \pi \left(a \mid \theta_0 + d\theta\right)^T \right]$$

So we start with:

$$\nabla_{d\theta} \log \pi \left(a \mid \theta_0 + d\theta\right) = \begin{bmatrix} \nabla_{\theta_\mu} \log \pi(a \mid \theta) \\ \nabla_{\theta_\sigma} \log \pi(a \mid \theta) \end{bmatrix} = \begin{bmatrix} \frac{a-\theta_\mu}{k^2\theta_\sigma^2} \\ -\frac{1}{\theta_\sigma} + \frac{(a-\theta_\mu)^2}{k^2\theta_\sigma^3} \end{bmatrix}$$

And the outer product looks as follows:

$$\nabla_{d\theta} \log \pi \left(a \mid \theta_0 + d\theta\right) \nabla_{d\theta} \log \pi \left(a \mid \theta_0 + d\theta\right)^T = \begin{bmatrix} \frac{(a-\theta_\mu)^2}{k^4\theta_\sigma^4} & \frac{(a-\theta_\mu)^3}{k^4\theta_\sigma^5} - \frac{(a-\theta_\mu)}{k^2\theta_\sigma^3} \\ \frac{(a-\theta_\mu)^3}{k^4\theta_\sigma^5} - \frac{(a-\theta_\mu)}{k^2\theta_\sigma^3} & \frac{(a-\theta_\mu)^4}{k^4\theta_\sigma^6} - \frac{2(a-\theta_\mu)^2}{k^2\theta_\sigma^4} + \frac{1}{\theta_\sigma^2} \end{bmatrix}$$

Taking the full expectation over the trajectories will simplify to expectations over the actions:

$$F_\theta = \begin{bmatrix} \mathbb{E}_a \frac{(a-\theta_\mu)^2}{k^4\theta_\sigma^4} & \mathbb{E}_a\left(\frac{(a-\theta_\mu)^3}{k^4\theta_\sigma^5} - \frac{(a-\theta_\mu)}{k^2\theta_\sigma^3}\right) \\ \mathbb{E}_a\left(\frac{(a-\theta_\mu)^3}{k^4\theta_\sigma^5} - \frac{(a-\theta_\mu)}{k^2\theta_\sigma^3}\right) & \mathbb{E}_a\left(\frac{(a-\theta_\mu)^4}{k^4\theta_\sigma^6} - \frac{2(a-\theta_\mu)^2}{k^2\theta_\sigma^4} + \frac{1}{\theta_\sigma^2}\right) \end{bmatrix}$$
$$= \begin{bmatrix} \frac{k^2\theta_\sigma^2}{k^4\theta_\sigma^4} & \frac{0}{k^4\theta_\sigma^5} - \frac{0}{k^2\theta_\sigma^3} \\ \frac{0}{k^4\theta_\sigma^5} - \frac{0}{k^2\theta_\sigma^3} & \frac{3k^4\theta_\sigma^4}{k^4\theta_\sigma^6} - \frac{2k^2\theta_\sigma^2}{k^2\theta_\sigma^4} + \frac{1}{\theta_\sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{k^2\theta_\sigma^2} & 0 \\ 0 & \frac{2}{\theta_\sigma^2} \end{bmatrix}$$

The second step in the equation above can be made by using the first to fourth moments of a Gaussian distribution, looking as follows:

$$\mathbb{E}_a(a-\theta_\mu) = 0$$
$$\mathbb{E}_a(a-\theta_\mu)^2 = \sigma(\theta_\sigma)^2 = k^2\theta_\sigma^2$$
$$\mathbb{E}_a(a-\theta_\mu)^3 = 0$$
$$\mathbb{E}_a(a-\theta_\mu)^4 = 3\sigma(\theta_\sigma)^4 = 3k^4\theta_\sigma^4$$

3. Gradient update for vanilla policy gradient:

$$\theta_{t+1} = \theta_t + \alpha\nabla_{\theta_t} J(\theta_t)$$

And for natural policy gradient:

$$\theta_{t+1} = \theta_t + \alpha F_{\theta_t}^{-1}\nabla_{\theta_t} J(\theta_t)$$

Where $F_{\theta_t}^{-1}$ can be obtained by taking the inverse of the results of the previous question:

$$F_{\theta_t}^{-1} = \begin{bmatrix} \frac{1}{k^2\theta_\sigma^2} & 0 \\ 0 & \frac{2}{\theta_\sigma^2} \end{bmatrix}^{-1} = \begin{bmatrix} k^2\theta_\sigma^2 & 0 \\ 0 & \frac{\theta_\sigma^2}{2} \end{bmatrix}$$

And $\nabla_{\theta_t} J(\theta_t)$ can be obtained by calculating $J(\theta_t) = E_a(r) = E_a(a-a^2) = E_a(a) - E_a(a^2) = \theta_\mu - \left(\sigma(\theta_\sigma)^2 + E_a(a)^2\right) = \theta_\mu - k^2\theta_\sigma^2 - \theta_\mu^2$ and taking the derivative wrt both θ_μ and θ_σ:

$$\nabla_{\theta_t} J(\theta_t) = \begin{bmatrix} \nabla_{\theta_\mu} J(\theta_t) \\ \nabla_{\theta_\sigma} J(\theta_t) \end{bmatrix} = \begin{bmatrix} 1-2\theta_\mu \\ -2k^2\theta_\sigma \end{bmatrix}$$

Updating rule for vanilla and natural policy gradient:

$$\text{Vanilla}: \begin{bmatrix} \theta_{\mu,t+1} \\ \theta_{\sigma,t+1} \end{bmatrix} = \begin{bmatrix} \theta_{\mu,t} \\ \theta_{\sigma,t} \end{bmatrix} + \alpha \begin{bmatrix} 1-2\theta_\mu \\ -2k^2\theta_\sigma \end{bmatrix}$$
$$\text{Natural}: \begin{bmatrix} \theta_{\mu,t+1} \\ \theta_{\sigma,t+1} \end{bmatrix} = \begin{bmatrix} \theta_{\mu,t} \\ \theta_{\sigma,t} \end{bmatrix} + \alpha \begin{bmatrix} k^2\theta_\sigma^2 & 0 \\ 0 & \frac{\theta_\sigma^2}{2} \end{bmatrix} \begin{bmatrix} 1-2\theta_\mu \\ -2k^2\theta_\sigma \end{bmatrix}$$
$$= \begin{bmatrix} \theta_{\mu,t} \\ \theta_{\sigma,t} \end{bmatrix} + \alpha \begin{bmatrix} k^2\theta_\sigma^2 - 2k^2\theta_\sigma^2\theta_\mu \\ -k^2\theta_\sigma^3 \end{bmatrix}$$

Filling in the two parameter settings gives the following updates for a) and b) respectively:

$$\text{Vanilla}: \begin{bmatrix} \theta_{\mu,t+1} \\ \theta_{\sigma,t+1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0.98 \end{bmatrix} \text{ and } \begin{bmatrix} 1 \\ -1.99 \end{bmatrix}$$
$$\text{Natural}: \begin{bmatrix} \theta_{\mu,t+1} \\ \theta_{\sigma,t+1} \end{bmatrix} = \begin{bmatrix} 0.01 \\ 0.99 \end{bmatrix} \text{ and } \begin{bmatrix} 0.01 \\ 0.0099 \end{bmatrix}$$

Where natural policy gradient results in exactly equal values for the mean and variance of new policies ($\mu = \theta_\mu$ and $\sigma^2 = (k\theta_\sigma)^2$). We can see that a vanilla policy gradient would also result in equal

means for both parameter settings, but the parameter θ_σ becomes lower than 0 in one specification (which is not possible for the variance, but is caused by a weakly specified linear standard deviation).

The natural policy gradient updates the parameters more gently than the vanilla policy gradients does, it therefore also seems to handle weakly specified linear variance/std.

4. The Fisher Information matrix regulates the size of updates, making sure they will not be too radical. It will decrease the size an update step of θ_μ by dividing it by 100 if $k = 0.1$, it will increase the update size when $k = 10$. This mechanism works because of the diagonal structure of the matrix, taking the inverse simply results in an extra weight for the update step. The hyperparameter k works in combination with θ_σ and is therefore always making sure the updates will not be too small/large and convergence is reached.

5. As the question above (and below) indicate, there will obviously be a difference for vanilla and natural policy gradients. As natural policy gradient makes the gradient independent of parameters by use of the Fisher Information matrix, the graphs for all three parameter settings will look equal (as can be seen in). The plots of vanilla policy gradient do differ for different parameter settings as the reweighting of gradients is not performed.

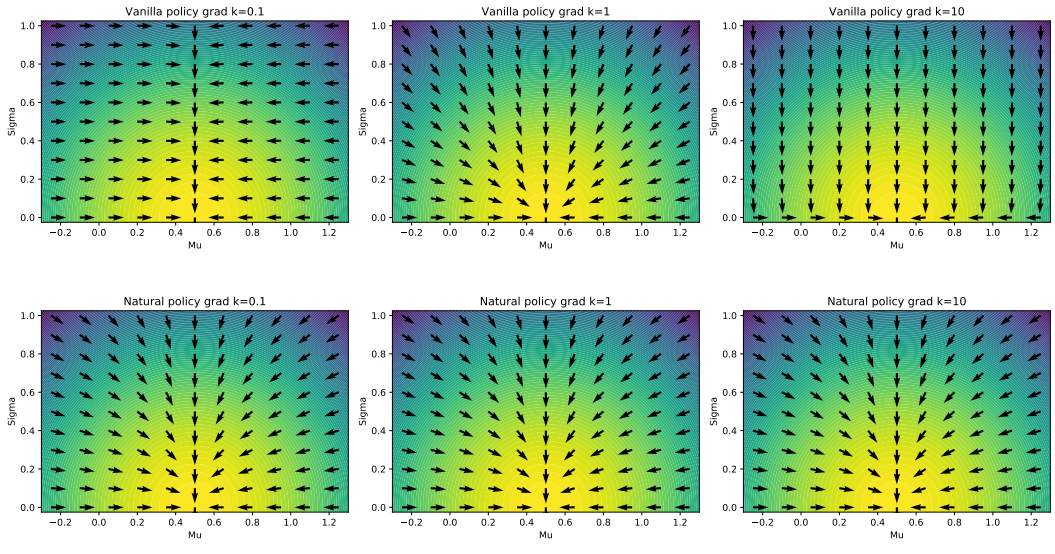

Figure 1: Gradient plots

6. The simultaneous updating of multiple parameters might lead to sub-optimal updating because the size of one of the gradients vastly exceeds the size of another gradient. It might be impossible to scale the updating steps by a single learning rate because the order of magnitude of the gradients is simply too different. Natural policy gradients will scale these differences down by "weighting" the specific gradient and making the updating independent of parameter size, this ensures updating steps in an improving direction. Natural policy gradient therefore has a higher probability of updating in the right direction and the concluding policy is expected to perform better (or at least faster).