RL-Week-8

Practice Assignment

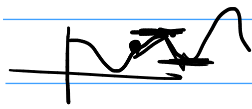
1) Which of the following is the correct way to represent policy for policy search methods? Assume θ_i **1 point**

- [x] $\pi(a|s) = \frac{e^{\theta_{as}}}{\sum_i e^{\theta_{is}}}$
- [ ] $\pi(a|s) = \frac{\theta_a}{\sum_i e^{\theta_i}}$
- [ ] $\pi(a|s) = \frac{e^{\theta_a}}{\sum_i \theta_i}$
- [ ] $\pi(a|s) = \frac{\theta_a^e}{\sum_i \theta_i^e}$

$$\pi(a|s) = \frac{e^{\theta_{as}}}{\sum_{i=0}^{\infty} e^{\theta_{is}}}$$

2) Suppose we are using policy gradient method to solve a reinforcement learning problem. Assuming that the policy returned by the method is not optimal, which among the following are plausible reasons for such an outcome? **1 point**

gradient ⇒ local maxima ⇒ no



- [x] The search procedure converged to a locally optimal policy.
- [x] The search procedure was terminated before it could reach an optimal policy.
- [x] An optimal policy could not be represented by the parameterisation used to represent the policy.
- [ ] None of thes

$\pi(w_t)$

3) Which of the following is the correct update rule for θ (representing the policy) with the policy gradient method? **1 point**

- [x] $\theta \leftarrow \theta + \alpha\nabla J(\theta)$
- [ ] $\theta \leftarrow \theta - \alpha\nabla J(\theta)$
- [ ] $\theta \leftarrow \theta + \nabla J(\theta)$
- [ ] $\theta \leftarrow \theta - \nabla J(\theta)$
- [ ] None of these

gradient ascent ↓ we move in the direcⁿ of the steepest ascent.

Max. $J(\theta) = E(r_t)$ → performance obj. funcⁿ for policy π.

θ updaⁿ rule ⇒ $\theta \leftarrow \theta + \alpha\nabla J(\theta)$

4) Which of the following is the correct formulation of cost function for multi arm bandit problem? **1 point**

- [x] $J(\theta) = \sum_a q^*(a)\pi_\theta(a)$
- [ ] $J(\theta) = \sum_a q(a)\pi_\theta(a)$
- [ ] $J(\theta) = q^*(a)\pi_\theta(a)$
- [ ] $J(\theta) = \sum_a q^*(a)\pi(a)$

$$J(\theta) = E(r_t) = \sum_a \underbrace{q_*(a)}_{\text{true expected reward}} \pi_\theta(a)$$

↓ mean of the gaussian.

5) Which of the following is the correct formulation to estimate cost function for multi arm bandit problem from N samples? r_i is the reward received after pulling an arm at i^{th} timestamp. **1 point**

- [x] $\hat\nabla J(\theta) = \frac{1}{N}\sum_{i=1}^N r_i \frac{\nabla\pi_a(\theta)}{\pi_a(\theta)}$
- [ ] $\hat\nabla J(\theta) = \frac{1}{N}\sum_{i=1}^N r_i \nabla\pi_a(\theta)$
- [ ] $\hat\nabla J(\theta) = \sum_{i=1}^N r_i \frac{\nabla\pi_a(\theta)}{\pi_a(\theta)}$
- [ ] $\hat\nabla J(\theta) = \frac{1}{N}\sum_{i=1}^N \frac{\nabla\pi_a(\theta)}{\pi_a(\theta)}$
- [ ] $\hat\nabla J(\theta) = \frac{1}{N}\sum_{i=1}^N r_i \frac{\pi_a(\theta)}{\nabla\pi_a(\theta)}$

avg.

$$\hat\nabla J(\theta) = \frac{1}{N}\sum_{i=1}^{N} r_i \underbrace{\frac{\nabla\pi_\theta(a_i)}{\pi_\theta(a_i)}}$$

c/d the likelihood ratio.

kind of importance sampling.

Graded Assignment

1) What are the advantages of policy search methods over other approaches? **1 point**

- [x] They can lead to simpler solution description.
- [x] They offer better convergence as compared to function approximation based methods.
- [x] In continuous action setting, they work better than value function based approaches.
- [x] They are robust to partial observability.
- [ ] None of the above

$\rho \Rightarrow$ probability $\rightarrow 0 \leq p \leq 1$

$\rightarrow$ addition should be always be 1

2) Which of the following is the correct way to represent policy for policy search methods? **1 point**

- [x] $\pi(a|s) = \rho_a, \sum_i \rho_i = 1$ and $1 \geq \rho_i \geq 0 \forall i$
- $\pi(a|s) = \rho_a, \sum_i \rho_i \geq 1$ and $1 \geq \rho_i \geq 0 \forall i$
- $\pi(a|s) = \rho_a, \sum_i \rho_i \leq 1$ and $1 \geq \rho_i \geq 0 \forall i$
- $\pi(a|s) = \rho_a, \sum_i \rho_i = 1$ and $2 \geq \rho_i \geq 0 \forall i$

$\rightarrow$ addition of probabilities of taking these action will always be one.

3) Which of the following is the correct update rule for policy parameter θ if the policy is represented in soft-max fashion for a multi arm bandit? **1 point**

- $\Delta\theta_i = \alpha(r - \bar{r})(1 - \pi(a_i, \theta))$
- $\Delta\theta_i = \alpha(r - \bar{r})(-\pi(a_i, \theta))$
- [x] $\Delta\theta_i = \begin{cases} \alpha(r - \bar{r})(1 - \pi(a_i, \theta)), \text{if action } a_i \text{ is chosen.} \\ \alpha(r - \bar{r})(-\pi(a_i, \theta)), \text{otherwise.} \end{cases}$
- None of these.

4) Consider following update rule of policy parameter (θ): $\Delta\theta_t = \alpha(r_t - b_t)\frac{\partial ln\pi_\theta(a_t)}{\partial\theta}$ **1 point**

r_t reinforcement

b_t baseline

Choose the correct statements of the following:

- [x] Baseline b_t can help figure out if the reward is high or low.
- [x] Baseline b_t can be set as the average of the rewards received.
- [x] $\frac{\partial ln\pi_\theta(a_t)}{\partial\theta}$ is characteristic eligibility, that decides how much a parameter gets updated.
- [ ] None of these.

5) Policy gradient methods can be used for continuous action spaces. **1 point**

- [x] True $\rightarrow$ always better choice than value based func^n for continuous action spaces.
- False