Reinforcement Learning HW4

Yonatan Ariel Slutzky Eyal Grinberg

17 July 2023

1 Theoretical Questions

1.1 Question 1

1.1.1 Section a

Let $t \in [T] \setminus \{1, 2\}$. Thus:

$$P(I_t = 2 \wedge \hat{\mu}_{2,T_2(t)} > \frac{\mu_a + \mu_b}{2}) \leq P(\hat{\mu}_{2,T_2(t)} > \frac{\mu_a + \mu_b}{2}) =$$

$$= P(\hat{\mu}_{2,T_2(t)} - \mu_a > \frac{\mu_b - \mu_a}{2})$$

Hence, The following holds from Hoeffding's inequality:

$$\sum_{t=3}^{T} P(I_t = 2 \wedge \hat{\mu}_{2,T_2(t)} > \frac{\mu_a + \mu_b}{2}) \leq \sum_{t=1}^{T} P(\hat{\mu}_{2,T_2(t)} - \mu_a > \frac{\mu_b - \mu_a}{2}) \leq$$

$$\leq \sum_{k=1}^{T} P(\hat{\mu}_{2,k} - \mu_a > \frac{\mu_b - \mu_a}{2}) \leq \sum_{k=1}^{\infty} P(\hat{\mu}_{2,k} - \mu_a > \frac{\mu_b - \mu_a}{2}) \leq$$

$$\sum_{k=1}^{\infty} exp(-\frac{2k^2(\mu_b - \mu_a)^2}{2^2 k}) = \sum_{k=1}^{\infty} (exp(-\frac{\Delta^2}{2}))^k = \frac{exp(-\frac{\Delta^2}{2})}{1 - exp(-\frac{\Delta^2}{2})} =$$

$$= \frac{1}{exp(\frac{\Delta^2}{2}) - 1} \leq \frac{2}{\Delta^2}$$

The above bound isn't dependant on T, as desired. □

1.1.2 Section b

Let $t \in [T] \setminus \{1, 2\}$. Thus:

$$P(I_t = 2 \wedge \hat{\mu}_{2,T_2(t)} \leq \frac{\mu_a + \mu_b}{2}) =$$

$$= P(\hat{\mu}_{2,T_2(t)} \leq \frac{\mu_a + \mu_b}{2} \wedge \hat{\mu}_{1,T_1(t)} \leq \frac{\mu_a + \mu_b}{2}) \cdot \chi_{t \in \mathbb{N}_{even}} \leq$$

$$\leq P(\hat{\mu}_{1,T_1(t)} - \mu_b \leq \frac{\mu_a - \mu_b}{2})$$

Hence, The following holds from Hoeffding's inequality:

$$\sum_{t=3}^{T} P(I_t = 2 \wedge \hat{\mu}_{2,T_2(t)} \leq \frac{\mu_a + \mu_b}{2}) \leq \sum_{t=1}^{T} P(\hat{\mu}_{1,T_1(t)} - \mu_b \leq \frac{\mu_a - \mu_b}{2}) \leq$$

$$\leq \sum_{k=1}^{T} P(\hat{\mu}_{1,k} - \mu_b \leq \frac{\mu_a - \mu_b}{2}) \leq \sum_{k=1}^{\infty} P(\hat{\mu}_{1,k} - \mu_b \leq \frac{\mu_a - \mu_b}{2}) \leq$$

$$\sum_{k=1}^{\infty} exp(-\frac{2k^2(\mu_b - \mu_a)^2}{2^2 k}) = \sum_{k=1}^{\infty}(exp(-\frac{\Delta^2}{2}))^k = \frac{exp(-\frac{\Delta^2}{2})}{1 - exp(-\frac{\Delta^2}{2})} =$$

$$= \frac{1}{exp(\frac{\Delta^2}{2}) - 1} \leq \frac{2}{\Delta^2}$$

The above bound isn't dependant on T, as desired. □

1.1.3 Section c

By definition and from our assumption and previous sections, the regret takes the following form:

$$Regret = E(max_{i \in [2]} \sum_{t=1}^{T} R_t(i) - \sum_{t=1}^{T} R_t(a_t)) =$$

$$= T\mu_b - \mu_b - \mu_a - \sum_{t=3}^{T} E(R_t(a_t)) = (T-2)\mu_b + \Delta - \sum_{t=3}^{T} \mu_b P(I_t = 1) + \mu_a P(I_t = 2) =$$

$$= \Delta + \sum_{t=3}^{T}(P(I_t = 1) + P(I_t = 2))\mu_b - \mu_b P(I_t = 1) - \mu_a P(I_t = 2) =$$

$$= \Delta + \sum_{t=3}^{T} \Delta P(I_t = 2) = \Delta + \Delta \sum_{t=3}^{T} P(I_t = 2) =$$

$$\Delta + \Delta \sum_{t=3}^{T} P(I_t = 2 \wedge \hat{\mu}_{2,T_2(t)} \leq \frac{\mu_a + \mu_b}{2}) + P(I_t = 2 \wedge \hat{\mu}_{2,T_2(t)} > \frac{\mu_a + \mu_b}{2}) \leq$$

$$\Delta + \Delta \cdot \frac{4}{\Delta^2} = \Delta + \frac{4}{\Delta}$$

The above bound isn't dependant on T as desired. □

1.2 Question 2

1.2.1 Section a

Given $\theta \in \mathbb{R}^d$ and $s \in S$, the probability of sampling within the interval $[0, 1]$ is

$$P(0 \leq a \leq 1|s;\theta) = P(0 \leq Exp(exp(\theta^\top \phi(s))) \leq 1) =$$

$$= F_{Exp(exp(\theta^\top \phi(s)))}(1) - F_{Exp(exp(\theta^\top \phi(s)))}(0) =$$

$$= 1 - exp(-exp(\theta^\top \phi(s)) \cdot 1) - (1 - exp(-exp(\theta^\top \phi(s)) \cdot 0)) =$$

$$= 1 - exp(-exp(\theta^\top \phi(s)))$$

The above is indeed a function of θ and s. □

1.2.2 Section b

For any $a \in \mathbb{R}_{<0}$, the density for a is 0 and thus the gradient w.r.t to θ is 0. Let $i \in [d]$. Thus, the following holds for any $a \in \mathbb{R}_{\geq 0}$:

$$\frac{\partial}{\partial \theta_i} log(\pi(a|s;\theta)) = \frac{\partial}{\partial \theta_i} log(exp(\theta^\top \phi(s)) exp(-exp(\theta^\top \phi(s)) \cdot a)) =$$

$$= \frac{\partial}{\partial \theta_i} log(exp(\theta^\top \phi(s) - exp(\theta^\top \phi(s)) \cdot a)) =$$

$$\frac{\partial}{\partial \theta_i} \theta^\top \phi(s) - exp(\theta^\top \phi(s)) \cdot a =$$

$$= \phi(s)_i - a \cdot \phi(s)_i exp(\theta^\top \phi(s)) = \phi(s)_i (1 - a \cdot exp(\theta^\top \phi(s)))$$

Thus, the gradient w.r.t to θ takes the following form

$$\nabla_\theta log(\pi(a|s;\theta) = (1 - a \cdot exp(\theta^\top \phi(s)))\phi(s)$$

as desired. □

1.2.3 Section c

We saw in class that the following holds for $J(\theta)$ which is the value function of the T-finite horizon:

$$\nabla J(\theta) \propto E^\pi(Q^\pi(s,a) \nabla_\theta log(\pi(a|s;\theta))) =: (*)$$

Plugging in the result from the previous section yields the following

$$(*) = E^\pi(Q^\pi(s,a)(1 - a \cdot exp(\theta^\top \phi(s)))\phi(s)) =$$

$$= E^\pi(Q^\pi(s,a))(1 - a \cdot exp(\theta^\top \phi(s)))\phi(s)$$

Thus, given U an unbiased estimator for $Q^\pi(s,a)$, the REINFORCE update takes the following form:

$$\Delta\theta = \alpha U (1 - a \cdot exp(\theta^\top \phi(s)))\phi(s)$$

where α is the learning rate used. □

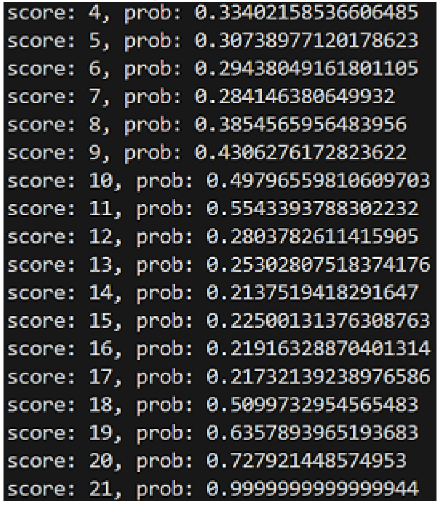
2 Practical Questions

The following questions taking into consideration the following dynamics:
- Multiple hits are possible for each player
- The MDP doesn't consider the open card of the dealer
- When getting to a hand whose sum is greater than 21, there's no point in hitting and thus the only available action for these states is to stand

In our modelling, the hit action is referred to as 1, and the stand action is referred to as 0.

2.1 Question 1

We ran the experiment with $\gamma = 1$ and with $lr = 0.01$ for $100,000$ iterations. The following are our results:



```
score: 4, prob: 0.33402158536606485
score: 5, prob: 0.30738977120178623
score: 6, prob: 0.29438049161801105
score: 7, prob: 0.284146380649932
score: 8, prob: 0.3854565956483956
score: 9, prob: 0.4306276172823622
score: 10, prob: 0.49796559810609703
score: 11, prob: 0.5543393788302232
score: 12, prob: 0.2803782611415905
score: 13, prob: 0.25302807518374176
score: 14, prob: 0.2137519418291647
score: 15, prob: 0.22500131376308763
score: 16, prob: 0.21916328870401314
score: 17, prob: 0.21732139238976586
score: 18, prob: 0.5099732954565483
score: 19, prob: 0.6357893965193683
score: 20, prob: 0.727921448574953
score: 21, prob: 0.9999999999999944
```

The results above are the probabilities for winning given we are in each of the states.

In order to compute the actual probability for winning, one would need to first compute the probability of reaching each of the states, which would be used when performing a weighted average of the results.

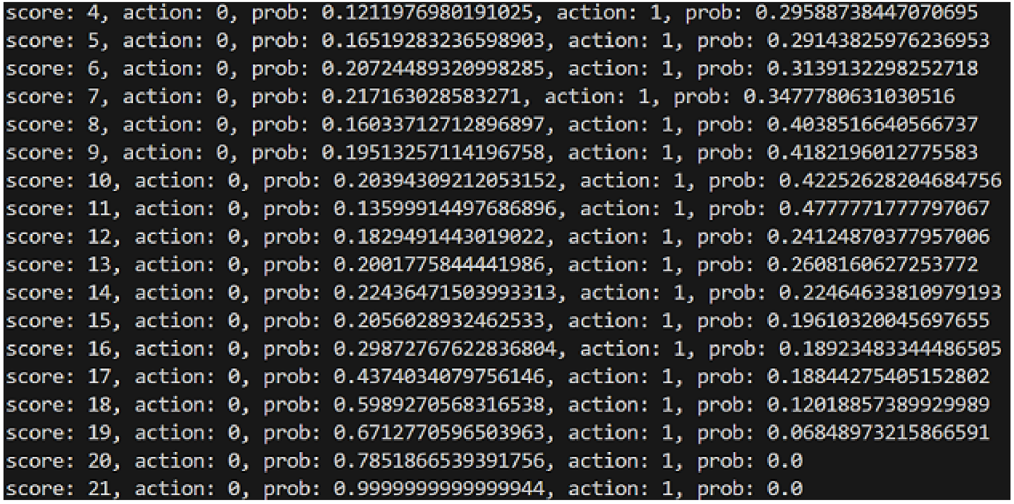
Then, in order to account for the dependence of the states, one would also need to apply the inclusion-exclusion principle which is a very tough task in our setting.

Thus, we used a surrogate, which instead of taking the probabilities of reaching each state, takes into account the probability of beginning in each of the states (this is much more easily computable and the events are now independent).

The estimated probability for winning we got after $100,000$ iterations using the proposed policy is 0.40498.

2.1.1 Question 2

We ran the experiment with $\gamma = 1$ and with $lr = 0.01$ for $100,000$ iterations. The following are our results:



```
score: 4, action: 0, prob: 0.1211976980191025, action: 1, prob: 0.29588738447070695
score: 5, action: 0, prob: 0.16519283236598903, action: 1, prob: 0.29143825976236953
score: 6, action: 0, prob: 0.20724489320998285, action: 1, prob: 0.3139132298252718
score: 7, action: 0, prob: 0.217163028583271, action: 1, prob: 0.3477780631030516
score: 8, action: 0, prob: 0.16033712712896897, action: 1, prob: 0.4038516640566737
score: 9, action: 0, prob: 0.19513257114196758, action: 1, prob: 0.4182196012775583
score: 10, action: 0, prob: 0.20394309212053152, action: 1, prob: 0.42252628204684756
score: 11, action: 0, prob: 0.13599914497686896, action: 1, prob: 0.4777771777797067
score: 12, action: 0, prob: 0.1829491443019022, action: 1, prob: 0.24124870377957006
score: 13, action: 0, prob: 0.2001775844441986, action: 1, prob: 0.2608160627253772
score: 14, action: 0, prob: 0.22436471503993313, action: 1, prob: 0.22464633810979193
score: 15, action: 0, prob: 0.2056028932462533, action: 1, prob: 0.19610320045697655
score: 16, action: 0, prob: 0.29872767622836804, action: 1, prob: 0.18923483344486505
score: 17, action: 0, prob: 0.4374034079756146, action: 1, prob: 0.18844275405152802
score: 18, action: 0, prob: 0.5989270568316538, action: 1, prob: 0.12018857389929989
score: 19, action: 0, prob: 0.6712770596503963, action: 1, prob: 0.06848973215866591
score: 20, action: 0, prob: 0.7851866539391756, action: 1, prob: 0.0
score: 21, action: 0, prob: 0.9999999999999944, action: 1, prob: 0.0
```

The optimal action in each state is the one with a higher probability of winning.
As expected, with higher states, the probability for winning when standing is increased, and the probability for winning when hitting is decreased.
An important insight (which is trivial for Blackjack players..) is that the highest probability for winning when hitting is achieved from state 11 (with a high probability we reach the higher states mentioned before).