Reinforcement Learning Assignment 4

Question 1:

a) We want to bound the expression $\sum_{t=1}^{T} \mathbb{P}[t \geq 3, I_t = 2, \hat{\mu}_{2,T_2(t)} > \frac{\mu_a - \mu_b}{2}]$.

Since $T_2(t) \leq t$ we can bound: $\mathbb{P}\left[t \geq 3, I_t = 2, \hat{\mu}_{2,T_2(t)} > \frac{\mu_a - \mu_b}{2}\right] \leq \mathbb{P}[\hat{\mu}_{2,t} > \frac{\mu_a - \mu_b}{2}]$ and by Hoeffding's inequality: $\mathbb{P}\left[\hat{\mu}_{2,t} > \frac{\mu_a - \mu_b}{2}\right] \leq \exp\,(-\frac{2t((\mu_a - \mu_b)^2)}{4})$ to get:

$\sum_{t=1}^{T} \mathbb{P}[t \geq 3, I_t = 2, \hat{\mu}_{2,T_2(t)} > \frac{\mu_a - \mu_b}{2}] \leq \sum_{t=1}^{T} \exp\,(-\frac{2t((\mu_a - \mu_b)^2)}{4})$

Since $\forall j, \exp(j) \geq 0$ we can use the hint and bound:

$$\sum_{t=1}^{T} \exp\,(-\frac{2t((\mu_a - \mu_b)^2)}{4}) \leq \sum_{t=1}^{\infty} \exp\,(-\frac{2t((\mu_a - \mu_b)^2)}{4})$$

The right-hand side expression is a geometric series, so we know that:

$$\sum_{t=1}^{\infty} \exp\,(-\frac{2t((\mu_a - \mu_b)^2)}{4}) = \frac{1}{\exp\left(\frac{(\mu_a - \mu_b)^2}{2}\right) - 1}$$

From here we will use the second hint $x + 1 \leq e^x$ to get:

$$\frac{1}{\exp\left(\frac{(\mu_a - \mu_b)^2}{2}\right) - 1} \leq \frac{2}{(\mu_a - \mu_b)^2}$$

So our final bound is: $\sum_{t=1}^{T} \mathbb{P}[t \geq 3, I_t = 2, \hat{\mu}_{2,T_2(t)} > \frac{\mu_a - \mu_b}{2}] \leq \frac{2}{(\mu_a - \mu_b)^2}$

b) We want to bound the expression $\sum_{t=1}^{T} \mathbb{P}[t \geq 3, I_t = 2, \hat{\mu}_{2,T_2(t)} \leq \frac{\mu_a - \mu_b}{2}]$.

Using the definition of the algorithm we know that:

$$\sum_{t=1}^{T} \mathbb{P}[t \geq 3, I_t = 2, \hat{\mu}_{2,T_2(t)} \leq \frac{\mu_a - \mu_b}{2}] \leq \sum_{t=3}^{T} \mathbb{P}[I_{t-1} = 1, \hat{\mu}_{1,T_1(t-1)} \leq \frac{\mu_a - \mu_b}{2}]$$

Again since $T_1(t) \leq t$ we can bound:

$\sum_{t=3}^{T} \mathbb{P}[I_{t-1} = 1, \hat{\mu}_{1,T_1(t-1)} \leq \frac{\mu_a - \mu_b}{2}] \leq \sum_{t=3}^{T} \mathbb{P}[\hat{\mu}_{1,t-1} \leq \frac{\mu_a - \mu_b}{2}] \leq \sum_{t=1}^{T} \mathbb{P}[\hat{\mu}_{1,t-1} \leq \frac{\mu_a - \mu_b}{2}]$

From here we can use Hoeffding again to bound:

$$\sum_{t=1}^{T} \mathbb{P}[\hat{\mu}_{1,t-1} \leq \frac{\mu_a - \mu_b}{2}] \leq \sum_{t=1}^{T} \exp\,(-\frac{2t((\mu_a - \mu_b)^2)}{4})$$

By the same calculation as before we get the final bound:

$$\sum_{t=1}^{T} \mathbb{P}[t \geq 3, I_t = 2, \hat{\mu}_{2,T_2(t)} \leq \frac{\mu_a - \mu_b}{2}] \leq \frac{2}{(\mu_a - \mu_b)^2}$$

c) Since $\Delta = \mu_a - \mu_b$, using our bounds from before together with the definition of regret we can bound:

$$\mathbb{E}[Regret] = \sum_{t=1}^{T} \Delta \mathbb{P}[I_t = 2] = \Delta + \sum_{t=3}^{T} \Delta \mathbb{P}[I_t = 2]$$

Using the union bound we get:

$$\sum_{t=3}^{T} \Delta \mathbb{P}[I_t = 2] \leq \Delta\left(\frac{2}{(\mu_a - \mu_b)^2} + \frac{2}{(\mu_a - \mu_b)^2}\right) = \frac{4}{\Delta}$$

Which gives an overall bound of $\Delta + \frac{4}{\Delta}$.

This bound does not depend on T which is exactly what we wanted.

Question 2:

a) The probability of sampling $a \in [0,1]$ is:

$$\int_0^1 \lambda \mathrm{e}^{-\lambda \mathrm{x}} dx = \left[-\frac{e^{-\lambda}}{\lambda}\right]_0^1 = 1 - \frac{e^{-\lambda}}{\lambda} = 1 - \exp\left(-\theta^T \phi(s)\right) \exp\left(-\exp\left(\theta^T \phi(s)\right)\right)$$

b) Calculating $\log(\pi(a|s,\theta)$ we get:

$$\log(\pi(a|s,\theta) = \log\left(\lambda \mathrm{e}^{-\lambda \mathrm{a}}\right) = \log(\lambda) - \lambda \mathrm{a}$$

From here we will use $\frac{d}{d\theta}\log(\lambda) = \phi(s)$ and $\frac{d\lambda}{d\theta} = \phi(s)\lambda$ to calculate:

$$\nabla_\theta \log(\pi(a|s,\theta) = \phi(s) - \phi(s)\lambda a = \phi(s) - \exp\left(\theta^T \phi(s)\right) \phi(s) a$$

c) The REINFORCEMENT update of the exponential distribution will be:

$$\theta^{t+1} = \theta^t + \alpha G \phi(s) D_\lambda(\theta^t)$$

Where G is the return and:

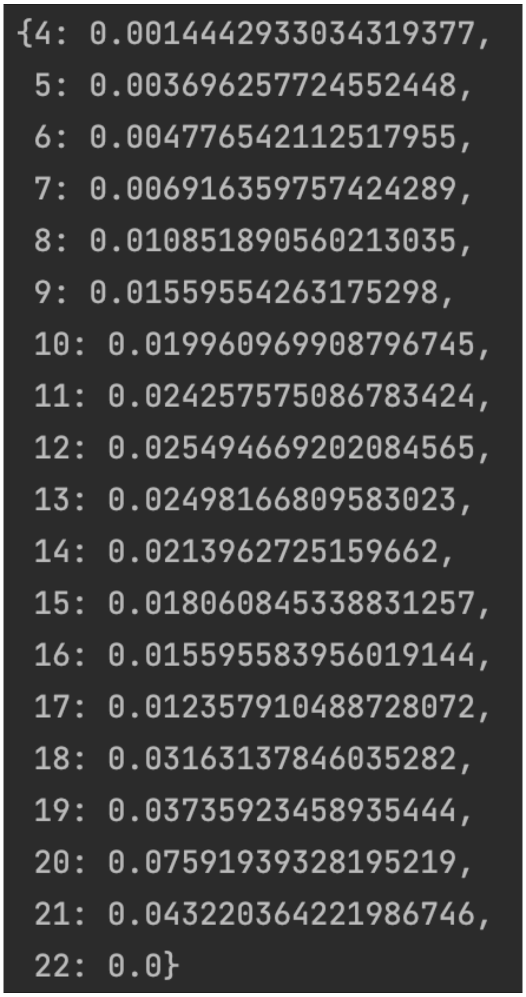
$$D_\lambda(\theta) = 1 - a \exp(\theta^T \phi(s))$$

Programming:

Q1:

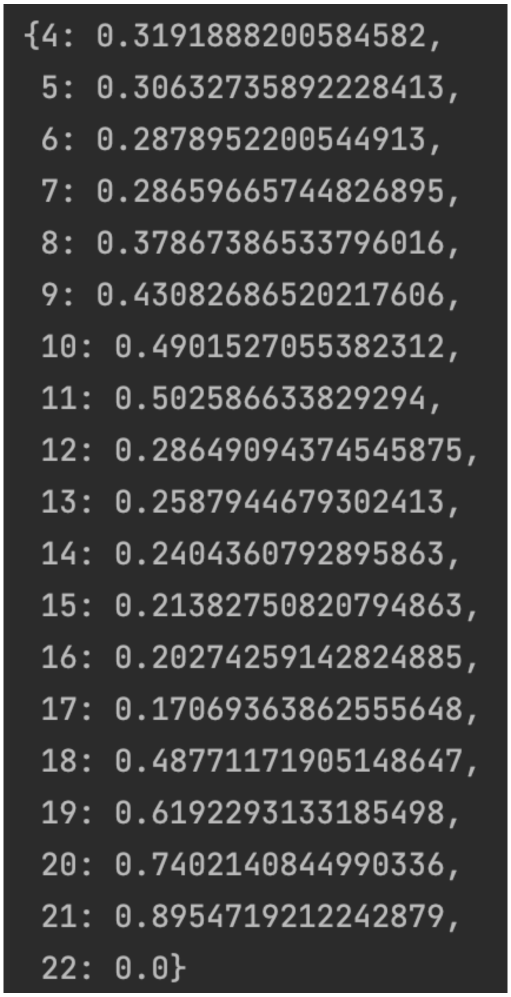
Attached are the results of running TD(0) in order to calculate the probabilities of the gambler winning from different states. Each state represents the sum of the two cards initially dealt. On the left are the joint probabilities: $P(outcome = winning, state = s)$ and on the right they are the conditional: $P(outcome = winning \mid state = s)$.
Also, the marginal probability $P(outcome = winning)$ is: 0.393.

$P(outcome = winning, state = s)$



```
{4: 0.0014442933034319377,
 5: 0.003696257724552448,
 6: 0.004776542112517955,
 7: 0.006916359757424289,
 8: 0.010851890560213035,
 9: 0.01559554263175298,
 10: 0.019960969908796745,
 11: 0.024257575086783424,
 12: 0.025494669202084565,
 13: 0.02498166809583023,
 14: 0.0213962725159662,
 15: 0.018060845338831257,
 16: 0.015595583956019144,
 17: 0.012357910488728072,
 18: 0.03163137846035282,
 19: 0.03735923458935444,
 20: 0.07591939328195219,
 21: 0.043220364221986746,
 22: 0.0}
```

$P(outcome = winning \mid state = s)$



```
{4: 0.3191888200584582,
 5: 0.30632735892228413,
 6: 0.2878952200544913,
 7: 0.28659665744826895,
 8: 0.37867386533796016,
 9: 0.43082686520217606,
 10: 0.4901527055382312,
 11: 0.502586633829294,
 12: 0.28649094374545875,
 13: 0.2587944679302413,
 14: 0.2404360792895863,
 15: 0.21382750820794863,
 16: 0.20274259142824885,
 17: 0.17069363862555648,
 18: 0.48771171905148647,
 19: 0.6192293133185498,
 20: 0.7402140844990336,
 21: 0.8954719212242879,
 22: 0.0}
```

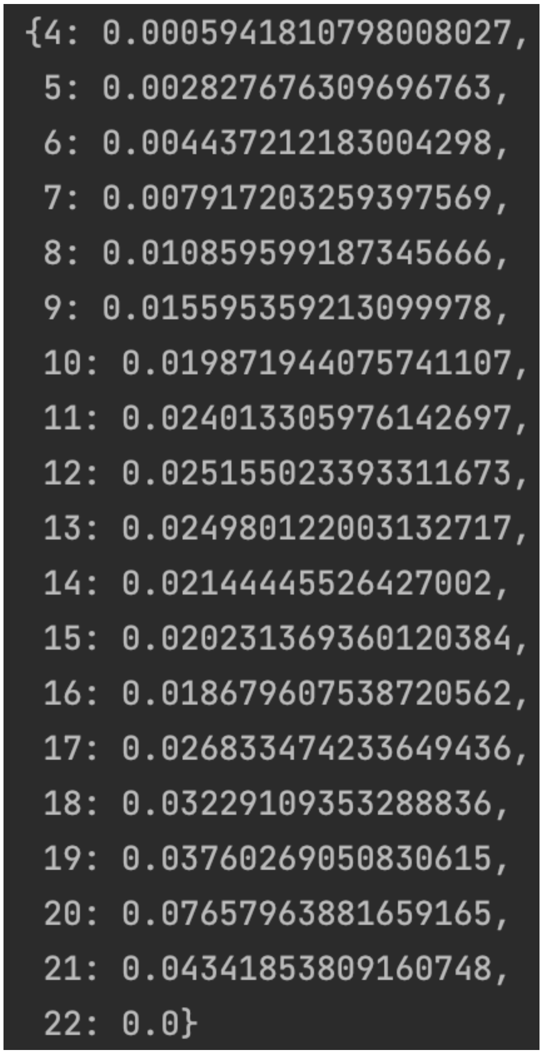
Q2:

Attached are the results of running SARSA in order to find the optimal policy for the gambler given the dealer's known policy. Attached are the optimal action a_{opt} for each state and the conditional probability: $P\left(outcome = winning \mid state = s, action = a_{opt}\right)$.

Also attached are the probabilities $P\left(outcome = winning,\ state = s \mid action = a_{opt}\right)$ which assume the gambler will pick a_{opt} as the action when reaching state s.

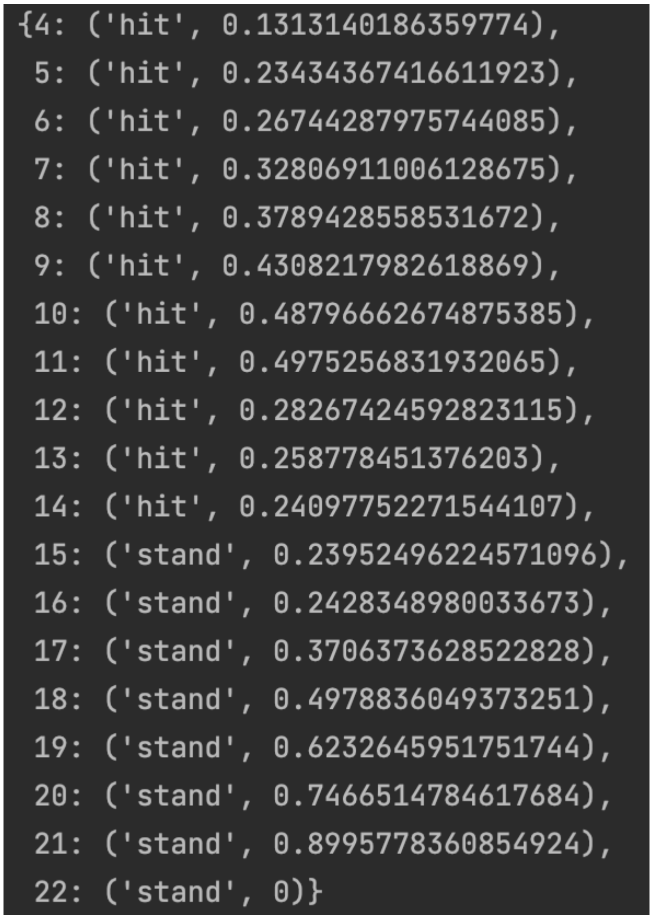
Also, the overall probability of winning if the actions chosen are the optimal ones is: $P\left(outcome = winning \mid action = a_{opt}\right)$ is: 0.413.

$P\left(outcome = winning, state = s \mid action = a_{opt}\right)$



```
{4: 0.0005941810798008027,
 5: 0.002827676309696763,
 6: 0.004437212183004298,
 7: 0.007917203259397569,
 8: 0.010859599187345666,
 9: 0.015595359213099978,
 10: 0.019871944075741107,
 11: 0.024013305976142697,
 12: 0.025155023393311673,
 13: 0.024980122003132717,
 14: 0.0214444552642700 2,
 15: 0.020231369360120384,
 16: 0.018679607538720562,
 17: 0.026833474233649436,
 18: 0.0322910935328883 6,
 19: 0.037602690508306 15,
 20: 0.076579638816591 65,
 21: 0.043418538091607 48,
 22: 0.0}
```

$P\left(outcome = winning \mid state = s, action = a_{opt}\right)$



```
{4: ('hit', 0.1313140186359774),
 5: ('hit', 0.23434367416611923),
 6: ('hit', 0.26744287975744085),
 7: ('hit', 0.3280691100612867 5),
 8: ('hit', 0.3789428558531672),
 9: ('hit', 0.4308217982618869),
 10: ('hit', 0.48796662674875385),
 11: ('hit', 0.4975256831932065),
 12: ('hit', 0.28267424592823115),
 13: ('hit', 0.258778451376203),
 14: ('hit', 0.24097752271544107),
 15: ('stand', 0.23952496224571096),
 16: ('stand', 0.2428348980033673),
 17: ('stand', 0.3706373628522828),
 18: ('stand', 0.4978836049373251),
 19: ('stand', 0.6232645951751744),
 20: ('stand', 0.7466514784617684),
 21: ('stand', 0.8995778360854924),
 22: ('stand', 0)}
```