A3

1. Rewawrd engineering

1.1

(a)

Consider a taxi-driver agent. Its reward would be composed of factors such as stability, duration, and whether it made it to the destination. The reward function seems fair, but the agent might still ignore the red light if the action doesn't decrease the reward.

(b)

`reward = helthy_reward + forward_reward - control_lost`

The goal of this environment would forward motion. It looks like being 'stable' might be an obstacle to an agent's learning, since motion already implies unbalance.

(c)

There are three conditions, and if any of them is satified, the Hopper is considered unhealthy, and the epsiode ends.

1. Elements of observation for the Hopper is above or under a healthy range. These elements include factors such as the angle, velocity, and angular velociy for each body part.
2. Hopper's height is no longer contained in the pre-defined closed interval. Which meaning the agent has fallen. (or flied)
3. The angle of the torso is no longer contained in the closed interval, meaning the agent has inclined too much to the front or back.

Early termination reduces the time steps required for meaningless learning, which leads to fast and efficient learning. However, since the agents do not experience the state after the fall (or the instable state) they will not be able to adjust to adversarial environments.

2. Learning from preferences

2.1

(a)

$$\nabla_w loss(\phi_w(o,a)) = \nabla_w \Big[- \sum_{(\sigma^1,\sigma^2,\mu)\in D} \mu(1)\log \hat{P}[\sigma^1 \succ \sigma^2] +$$

$$\mu(2)\log\hat{P}[\sigma^2 \succ \sigma^1]\Big]$$

$$\hat{P}[\sigma^1\succ\sigma^2] = \frac{\exp(\phi_w(\sigma^1))}{\exp(\phi_w(\sigma^1))+\exp(\phi_w(\sigma^2))} = \text{sigmoid}(\phi_w(\sigma^1) - \phi_w(\sigma^2)) = \Delta(\phi_{w,1})$$

$$\phi_w(\sigma) = \sum_{(o,a)\in\sigma}\phi_w(o,a)$$

$$= -\nabla_w \sum_D \Big[\mu(1)\log\sigma_{sig}(\Delta\phi_{w,1}) + \mu(2)\log\sigma_{sig}(\Delta\phi_{w,2})\Big]$$

$$\nabla \log\sigma(f) = f' \cdot \tfrac{\sigma'(f)}{\sigma(f)} = f'(1-\sigma(f))$$

$$= -\textstyle\sum_D \mu(1)\nabla_w(\Delta\phi_{w,1})(1-\sigma(\Delta\phi_{w,1})) + \mu(2)\nabla_w(\Delta\phi_{w,2})(1-\sigma(\Delta\phi_{w,2}))$$

$\Delta\phi_{w,1} = -\Delta\phi_{w,2}$, $1-\sigma(\Delta\phi_{w,2}) = \sigma(\Delta\phi_{w,1})$

$$= -\textstyle\sum_D \nabla_w(\Delta\phi_1)\Big[\mu(1)(1-\sigma(\Delta\phi_1) - (1-\mu(1))\sigma(\Delta\phi_1)\Big]$$

$$= -\textstyle\sum_D \nabla_w(\Delta\phi_1)\Big[\mu(1) - \sigma(\Delta\phi_{w,1})\Big]$$

$$\therefore \nabla_w L = -\sum_{(\sigma^1,\sigma^2)\in D}(\mu(1)-\hat{P}[\sigma^1\succ\sigma^2])(\sum_{(o,a)\in\sigma^1}\nabla_w\phi_w(o,a) - \sum_{(o,a)\in\sigma^2}\nabla_w\phi_w(o,a)$$

4. Best Arm Identification in Multi-armed Bandit

(a)

if $r_a^1, \dots, r_a^{n_e}$ are i.i.d. random variables satisfying $0 \leq r_a^i \leq 1$ with probability 1 for all i, $\bar{r}_a = \mathbb{E}[r_a^1] = \dots, = \mathbb{E}[r_a^{n_e}]$ is the expected value of the random variables, and $\hat{r}_a = \frac{1}{n}\sum_i^{n_e} r_a^i$ is the sample mean, then for any $\delta > 0$ we have

$$Pr\Big(|\hat{r}_a - \bar{r}_a| > \sqrt{\frac{\log(2/\delta)}{2n_e}}\Big) < \delta$$

$$Pr(\exists a \in A \ s.t. \ |\hat{r}_a - \bar{r}_a| > U)$$

: probability of our sample mean is not bounded for at least one action

E_a : event that $|\hat{r}_a - \bar{r}_a| > U$

$= Pr(\cup_{a \in A} E_a) \leq \sum_{a \in A} Pr(E_a) < \sum_{a \in A} \delta = |A|\delta$

$Pr(\exists a \in A \; s.t. \; |\hat{r}_a - \bar{r}_a| > U) < |A|\delta$

⇒ The more actions we can choose the more likely that our estimate can fail.

(b)

1. How degree of accuracy is $\bar{r}_{a^\dagger} \geq \bar{r}_{a^*} - \epsilon$?

$\hat{r}_{a^*} \leq \hat{r}_{a^\dagger} \; (a^\dagger = \arg\max_a \hat{r}_a)$

let say we have bounded

$|\hat{r}_a - \bar{r}_a| \leq u, \forall a$

then

$\hat{r}_{a^\dagger} \leq \bar{r}_{a^\dagger} + u$

$\hat{r}_{a^*} \geq \bar{r}_{a^*} - u$

combine these we get

$\bar{r}_{a^\dagger} + u \geq \hat{r}_{a^\dagger} \geq \hat{r}_{a^*} \geq \bar{r}_{a^*} - u$

$\rightarrow \bar{r}_{a^\dagger} \geq \bar{r}_{a^*} - 2u$

Our target accuracy is $|\hat{r}_a - \bar{r}_a| \leq \frac{\epsilon}{2}$

2. How much total samples (**accross all arms)** do wee need to return an ϵ optimal arm with prob more than $1 - \delta'$?

   ⇒ Hold $|\hat{r}_a - \bar{r}_a| > \frac{\epsilon}{2}$ with probability less than δ'

recall $Pr(\exists a \in A \; s.t. \; |\hat{r}_a - \bar{r}_a| > U) < |A|\delta$

$U = \frac{\epsilon}{2}, \delta = 2 \cdot \exp(-n_e \epsilon^2 / 2)$

we want $|A|\delta \leq \delta' \rightarrow |A| \cdot 2 \cdot \exp(-n_e \epsilon^2/2)$

$n_e \geq \frac{2}{\epsilon^2} \ln(\frac{2|A|}{\delta'})$

n_e : sampling number of each arm $(\hat{r}_a = \frac{1}{n_e} \sum_i^{n_e} r_a^i)$

$$N = |A| n_e \geq \frac{2|A|}{\epsilon^2} \ln(\frac{2|A|}{\delta'})$$

(c)

Target accuarcy (same) : $|\hat{r}_a - \bar{r}_a| \leq \frac{\epsilon}{2}$ prob. more than $1 - \delta$

$Pr(E_a) = Pr(\cup_{(1,2) \in A} |\hat{r}_a - \bar{r}_a| > \frac{\epsilon}{2}) \leq \delta$

$Pr(|\hat{r}_a - \bar{r}_a| > \frac{\epsilon}{2}) \leq \frac{\delta}{2}$

By CLT

$\hat{r}_a \sim \mathcal{N}(\bar{r}_a, \sigma_a^2/n_e)$

$$Z = \frac{\hat{r}_a - \bar{r}_a}{\sigma_a/\sqrt{n_e}}$$

$$Pr(|Z| > \frac{\epsilon/2}{\sigma_a/\sqrt{n_e}}) \leq \delta/2$$

$\vdots$

$$n_e \geq \frac{4\sigma^2}{\epsilon^2}(z_1 - \delta/4)^2$$

$N = 2n_e$

- Hoeffding $N = 2n_e \geq \dfrac{4}{\epsilon^2}\ln(\dfrac{4}{\delta'})$
  - robust : there is no assumpton on the distirbution
  - conservative bound
- Normal : $\dfrac{8\sigma^2}{\epsilon^2}(z_1 - \delta/4)^2$
  - efficient
  - depends on variance of the distribution