[Mustafa Suman] Assignment 5

Due November 28, 11:59 pm

1 Analysis

1.1

We look at our optimization goal:

Def. & using $\bar{r}$

$$\pi \overset{\downarrow}{=} \arg\max_{\pi} \mathbb{E}_{s,a\sim\rho_\pi}\left[Q(s,a) + \lambda\, b(s,a)\right]$$

$$= \arg\max_{\pi} \mathbb{E}_{s,a\sim\rho_\pi}\left[Q(s,a)\right] + \lambda \cdot \mathbb{E}_{s,a\sim\rho_\pi}\left[b(s,a)\right]$$

Considering $\left(\arg\max_{\pi} \mathbb{E}_{\mathbf{s},\mathbf{a}\sim p_\pi}[Q(\mathbf{s},\mathbf{a})] - \lambda D(\pi,\pi_\beta)\right) = \left(\arg\max_{\pi} \mathbb{E}_{\mathbf{s},\mathbf{a}\sim p_\pi}[Q(\mathbf{s},\mathbf{a})] \text{ s.t. } D(\pi,\pi_\beta) \le \varepsilon\right)$,

we derive:

$$\mathbb{E}_{s,a\sim\rho_\pi}\left[b(s,a)\right] \overset{!}{=} -D(\pi,\pi_\beta)$$

$$\overset{\text{target}\downarrow}{=} -\mathbb{E}_{s\sim\rho_\pi} D_{KL}\left[\pi(a|s)\,\|\,\pi_\beta(a|s)\right]$$

$$\overset{\text{Def.}}{=} -\mathbb{E}_{s\sim\rho_\pi}\mathbb{E}_{a\sim\rho_\pi}\left[\log\frac{\pi(a|s)}{\pi_\beta(a|s)}\right]$$

$$= \mathbb{E}_{s,a\sim\rho_\pi}\left[-\log\frac{\pi(a|s)}{\pi_\beta(a|s)}\right]$$

Hence, defining $\boxed{b(s,a) := -\log\frac{\pi(a|s)}{\pi_\beta(a|s)}}$, we can enforce our desired constraint.

1.2

Following the same approach, we derive:

$$\mathbb{E}_{s,a\sim\rho_\pi}[b(s,a)] = -\mathbb{E}_{s\sim\rho_\pi}\,\mathbb{E}_{\pi_\beta(a|s)}\left[f\left(\frac{\pi(a|s)}{\pi_\beta(a|s)}\right)\right]$$

$$= -\mathbb{E}_{s\sim\rho_\pi}\,\mathbb{E}_{\pi(a|s)}\left[f\left(\frac{\pi(a|s)}{\pi_\beta(a|s)}\right)\cdot\frac{\pi_\beta(a|s)}{\pi(a|s)}\right]$$

$$= \mathbb{E}_{s,a\sim\rho_\pi}\left[-f\left(\frac{\pi(a|s)}{\pi_\beta(a|s)}\right)\cdot\frac{\pi_\beta(a|s)}{\pi(a|s)}\right]$$

Hence, defining $b(s,a) := -f\left(\frac{\pi(a|s)}{\pi_\beta(a|s)}\right)\cdot\frac{\pi_\beta(a|s)}{\pi(a|s)}$ we can enforce our desired constraint.

1.3

Looking at the KL-Divergence between trajectories, we get

$$D_{KL}(p_\pi(\tau) \,\|\, p_{\pi_\beta}(\tau)) \overset{\text{Def.}}{=} \mathbb{E}_{\tau\sim p_\pi}\left[\log \frac{p_\pi(\tau)}{p_{\pi_\beta}(\tau)}\right]$$

$$\overset{\substack{\text{Def.}\\ \& \\ \text{Hint}}}{=} \mathbb{E}_{\tau\sim p_\pi}\left[\log \frac{\prod_t \int_{a\in\mathcal{A}} p(s_{t+1}|s_t,a)\cdot\pi(a|s_t)\,da}{\prod_t \int_{a\in\mathcal{A}} p(s_{t+1}|s_t,a)\cdot\pi_\beta(a|s_t)\,da}\right]$$

$$\overset{\substack{\log-\\ \text{Law}}}{=} \mathbb{E}_{\tau\sim p_\pi}\left[\sum_t \log \frac{\int_{a\in\mathcal{A}} p(s_{t+1}|s_t,a)\cdot\pi(a|s_t)\,da}{\int_{a\in\mathcal{A}} p(s_{t+1}|s_t,a)\cdot\pi_\beta(a|s_t)\,da}\right]$$

$$\overset{\text{Linearity}}{=} \sum_t \mathbb{E}_{\tau\sim p_\pi}\left[\log \frac{\int_{a\in\mathcal{A}} p(s_{t+1}|s_t,a)\cdot\pi(a|s_t)\,da}{\int_{a\in\mathcal{A}} p(s_{t+1}|s_t,a)\cdot\pi_\beta(a|s_t)\,da}\right]$$

$$= \mathbb{E}_{\substack{s_t,a,s_{t+1}\\ \sim p_\pi}}\left[\log \frac{\int_{a\in\mathcal{A}} p(s_{t+1}|s_t,a)\cdot\pi(a|s_t)\,da}{\int_{a\in\mathcal{A}} p(s_{t+1}|s_t,a)\cdot\pi_\beta(a|s_t)\,da}\right]$$

Now, defining $\boxed{b(s,a,s') := \log \frac{\int_{a\in\mathcal{A}} p(s'|s,a)\cdot\pi(a|s)\,da}{\int_{a\in\mathcal{A}} p(s'|s,a)\cdot\pi_\beta(a|s)\,da}}$

and $\boxed{b(s,a) = \mathbb{E}_{s'\sim p(\cdot|s,a)}\left[b(s,a,s')\right]}$, we get that

$$\mathbb{E}_{s,a\sim p_\pi}\left[b(s,a)\right] = D_{KL}(p_\pi(\tau) \,\|\, p_{\pi_\beta}(\tau))$$

enforces our desired constraint.

3 Exploration

3.1 Running a random policy

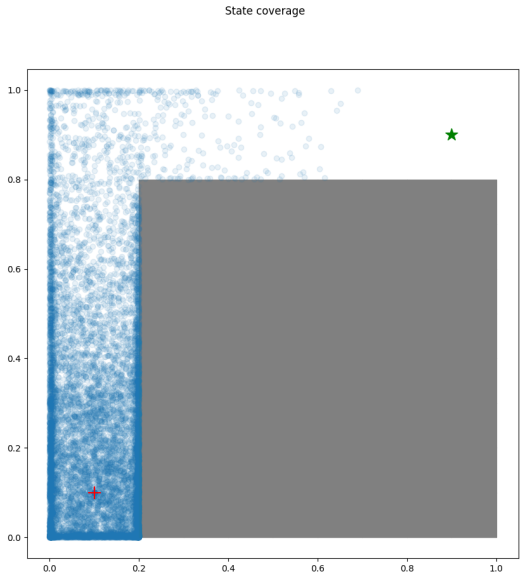

Figure 1: Random policy exploration; Environment: Easy

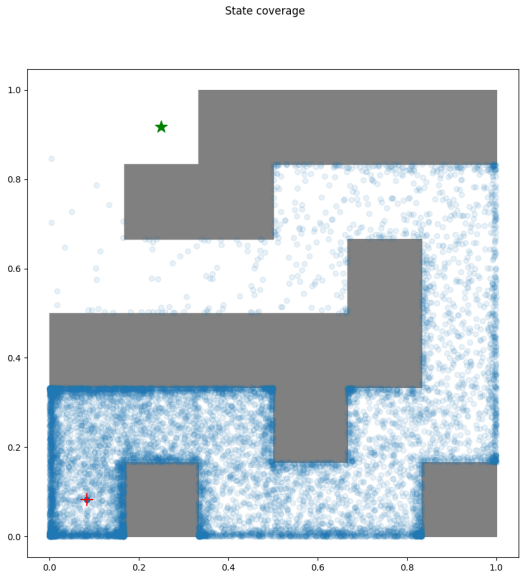

Figure 2: Random policy exploration; Environment: Medium

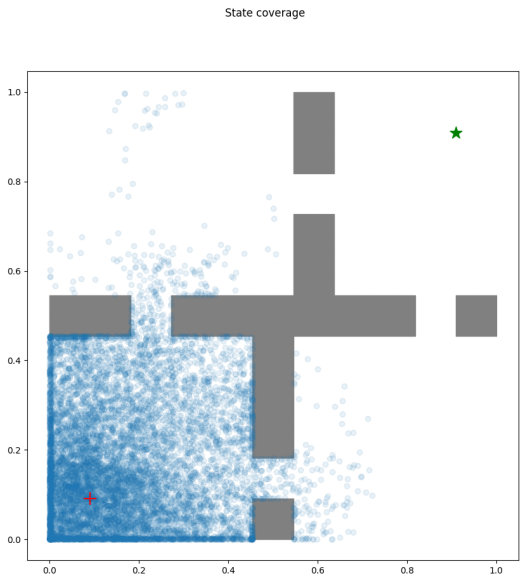

Figure 3: Random policy exploration; Environment: Hard

3.2 Random Network Distillation

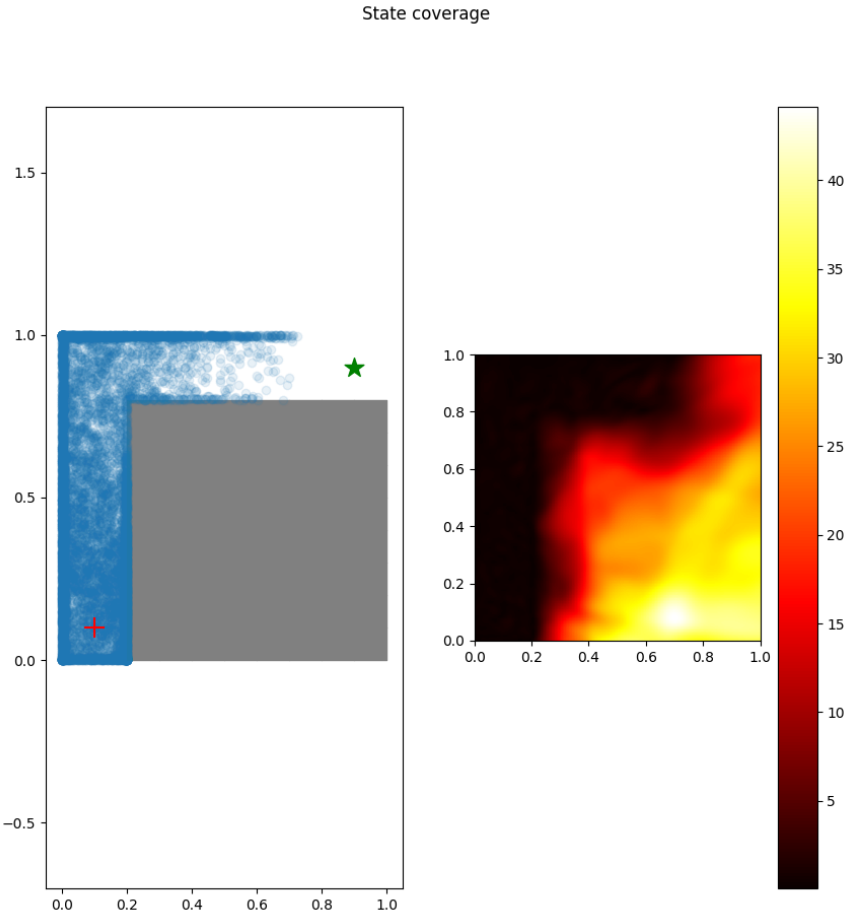

Figure 4: Random Network Distillation; Environment: Easy

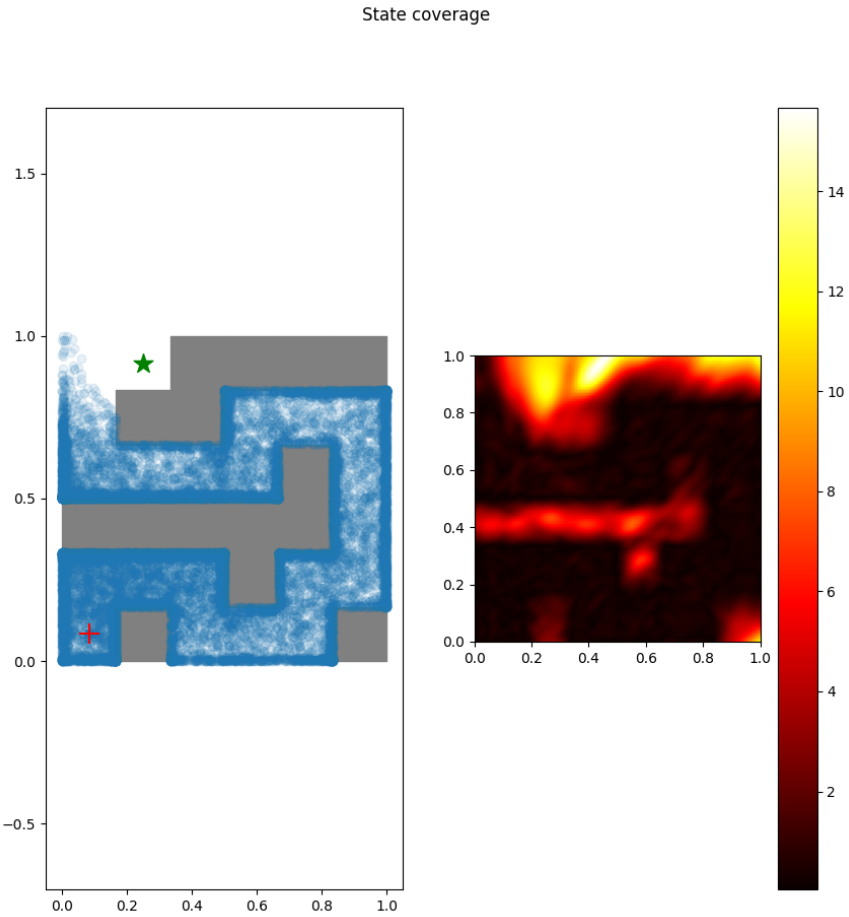

Figure 5: Random Network Distillation; Environment: Medium

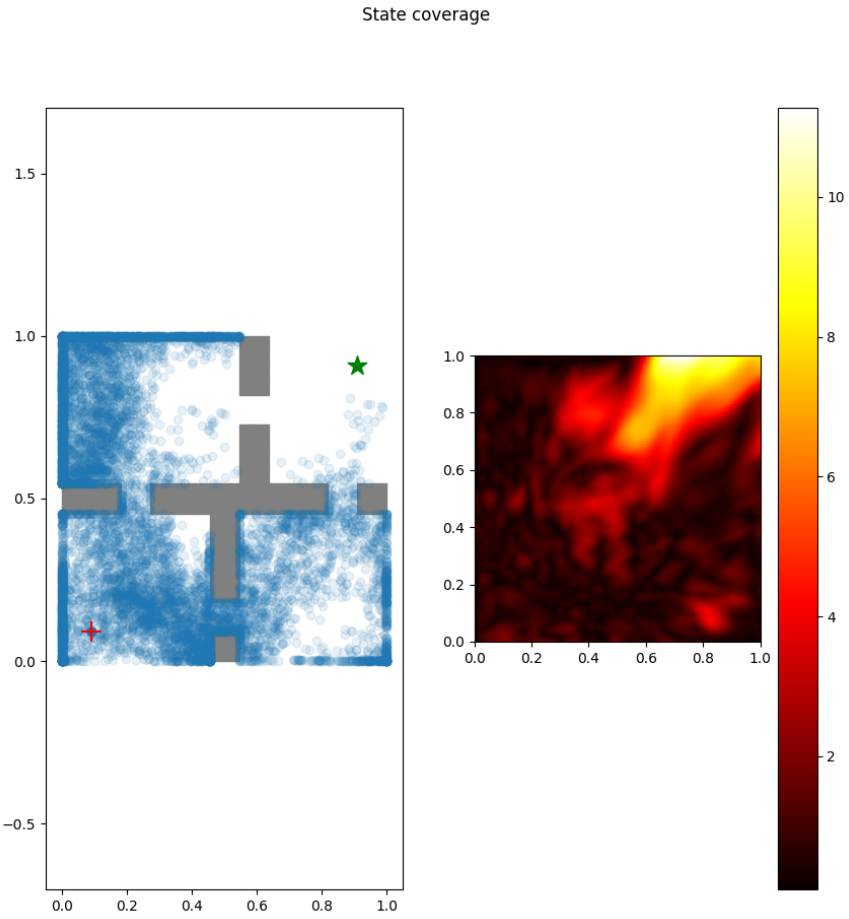

Figure 6: Random Network Distillation; Environment: Hard

4 Offline RL

4.1 CQL

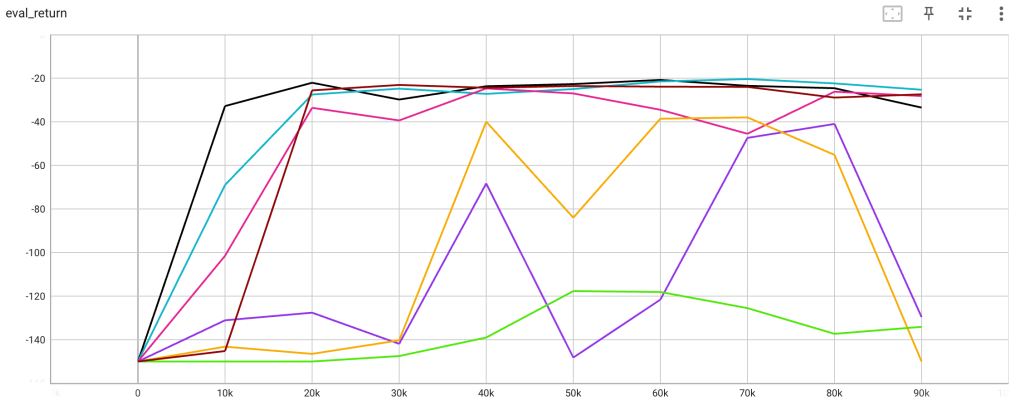

Figure 7: Eval return comparison for different alpha values in CQL setting. Black = 0.1, Blue = 0.2, Pink = 0.4, Orange = 0.6, Purple = 0.8, Green = 1, Dark Red = DQN; Environment: Medium

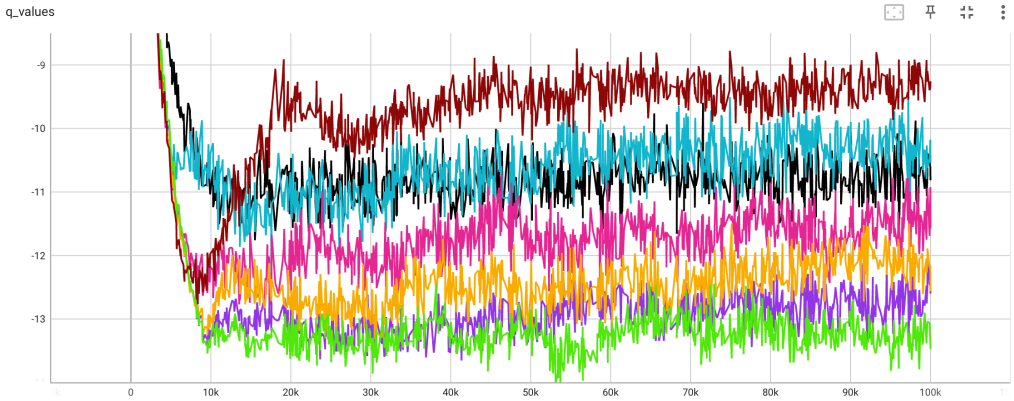

Figure 8: Q-Value comparison for different alpha values in CQL setting. Black = 0.1, Blue = 0.2, Pink = 0.4, Orange = 0.6, Purple = 0.8, Green = 1, Dark Red = DQN; Environment: Medium

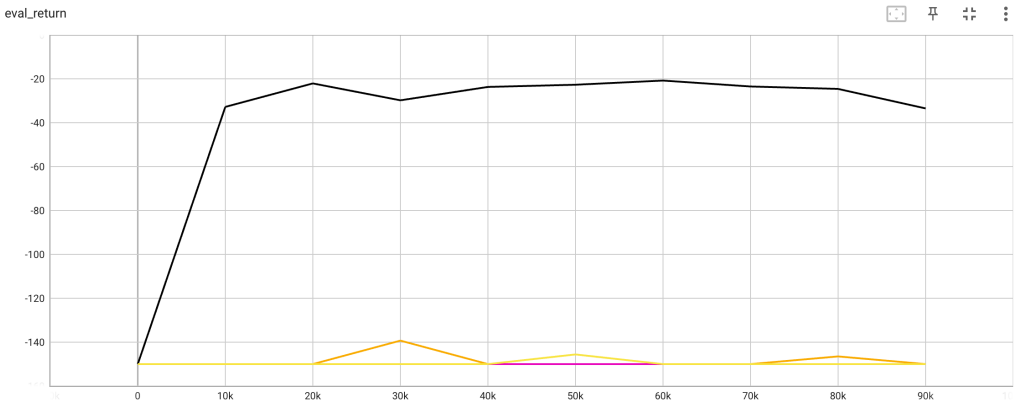

Figure 9: Eval return comparison for different alpha values in CQL setting. Black = 0.1, Yellow = 10, Green = 8, Pink = 6, Purple = 4, Orange = 2; Environment: Medium

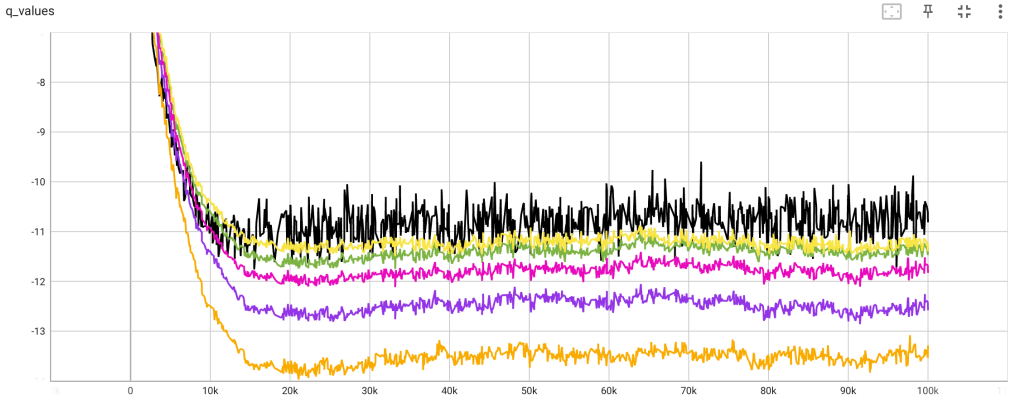

Figure 10: Q-Value comparison for different alpha values in CQL setting. Black = 0.1, Yellow = 10, Green = 8, Pink = 6, Purple = 4, Orange = 2; Environment: Medium

We can see that for an α which is greater than 1, the eval return is nearly a flat line representing that there is no real learning process. Therefore, we focus on the first two plots where the α value is in the range of 0 to 1. What we can see is that the best value in terms of success is 0.1 and the bigger α gets, the worse the algorithm performs. Considering the Q-Values, we can observe that the bigger α gets, the smaller the Q-Values become.

4.2 Policy Constraint Methods: IQL and AWAC

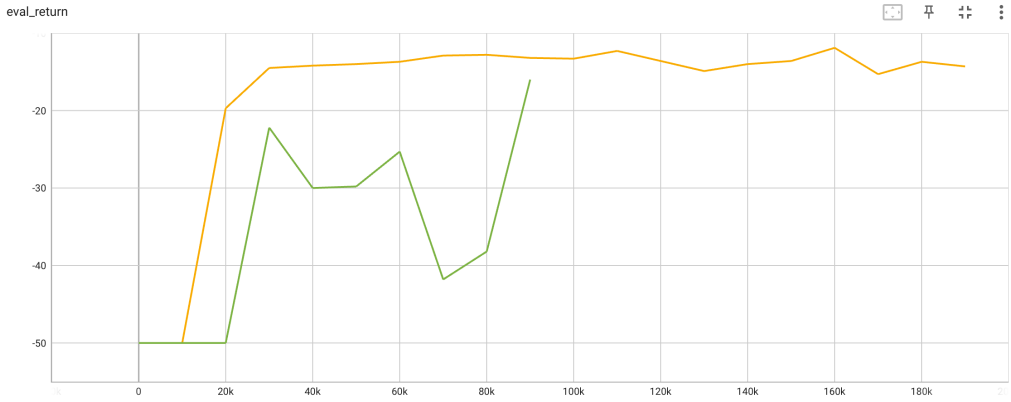

Figure 11: Eval return comparison for AWAC (=green) and IQL (=orange); Environment: Easy

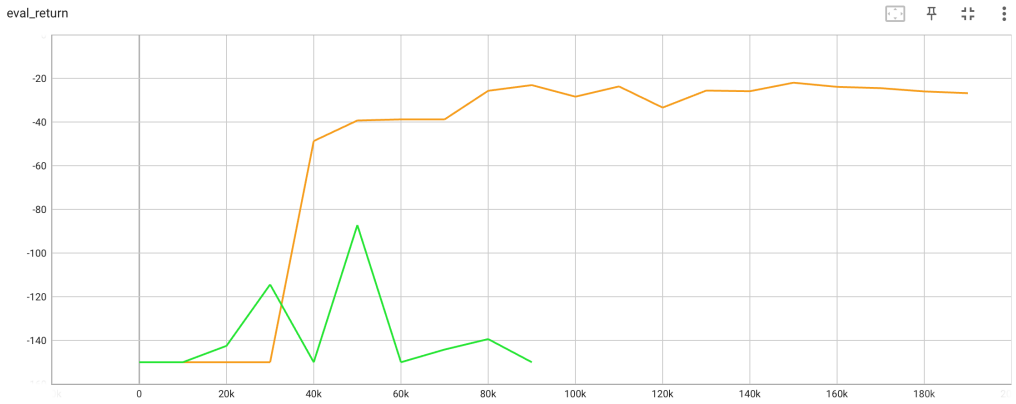

Figure 12: Eval return comparison for AWAC (=green) and IQL (=orange); Environment: Medium

4.3 Data Ablation - CQL

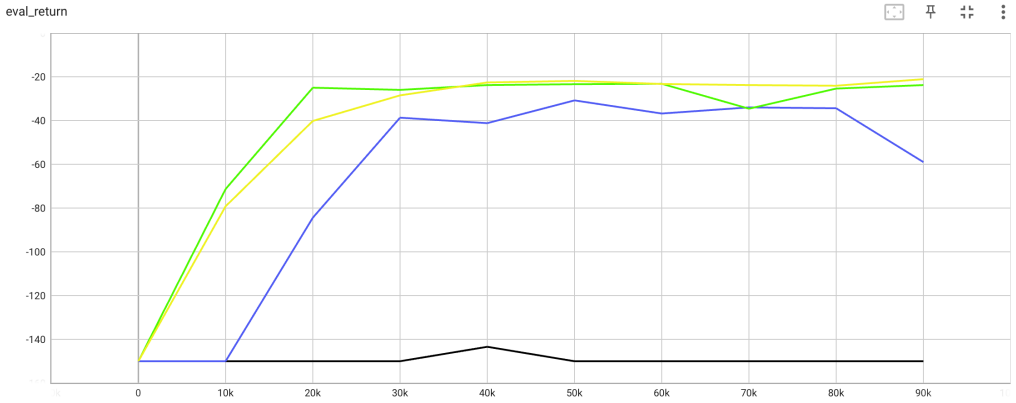

Figure 13: Eval return comparison for different sizes of exploration data sets in the CQL setting; Number of samples in each data set equals Black = 1K, Blue = 5K, Green = 10K and Yellow = 20K; Environment: Medium

We can see that the CQL Algorithm performs better for a bigger data set, which intuitively makes total sense. Due to our exploration concept, the more data we have, the more likely it gets that we have a higher state coverage and also observing high reward transitions. There is a ceiling though when we focus on the fact that the 10K and 20K processes perform nearly the same!

Note that for generating the plots, I simply made use of Tensorboard. So no special instructions are needed. The only thing I've changed (deviating from the defaults) is the discount factor to 0.98 in the AWAC Easy Environment.