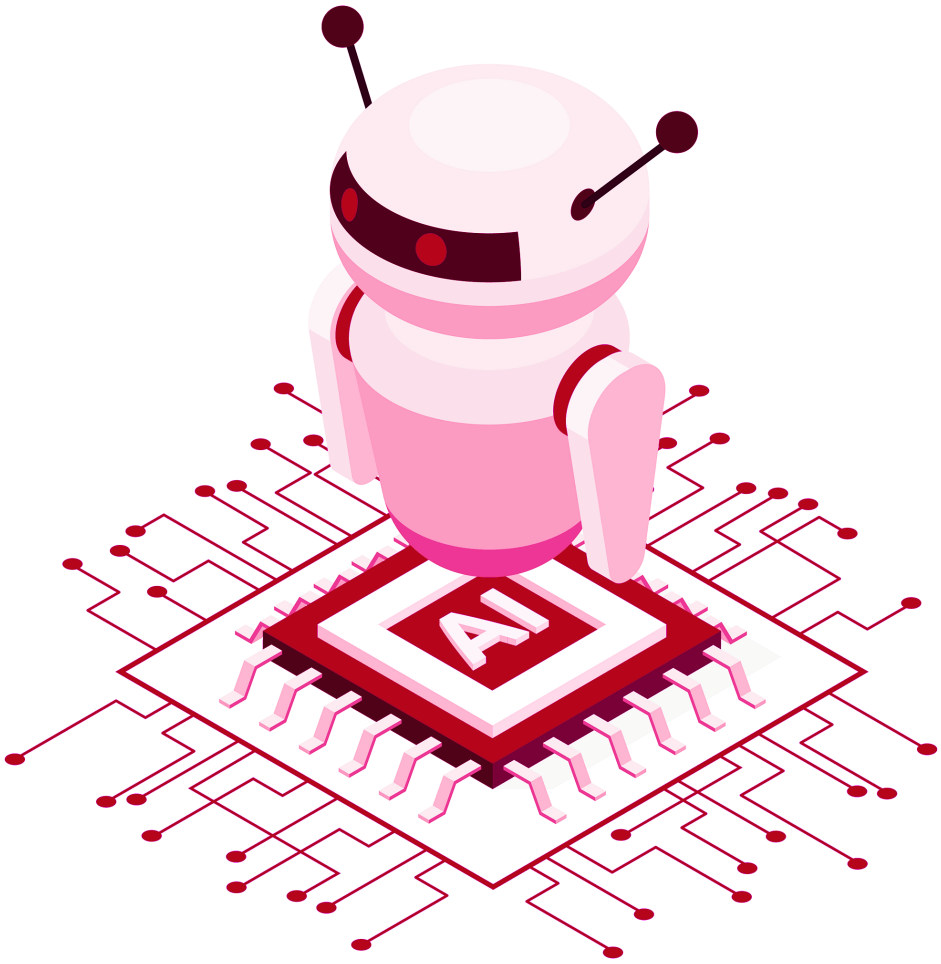

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution for Homework 12:

Offline Methods

By:

[Full Name]

[Student Number]

RIML

Deep RL Course

Spring 2025

Contents

1 Part 1 [60-points]

1. Considering the Bellman update, explain with reasoning why value estimation suffers from overestimation in the offline framework. [10-points]

$$Q(s,a) \leftarrow r(s,a) + \mathbb{E}_{a' \sim r_{new}}[Q(s',a')]$$

Answer: It is due to a phenomenon known as distributional shift. The learning agent is trained on a static dataset collected by a behavior policy, but it tries to learn an optimal policy (π new) that may be very different. The Bellman update equation, , involves estimating the value of taking actions in the next state according to this new policy.

The core problem arises when the new policy selects actions that are out-of-distribution. Actions that were rarely or never taken in the original dataset for a given state. Since the Q-function approximator has not been trained on these OOD state-action pairs, its value estimates for them are unreliable and can be arbitrarily high due to extrapolation errors.

The maximization inherent in the Bellman update will exploit these erroneously high Q-values. This error then gets propagated backward to other state-action values during training, leading to a cascading overestimation across the entire value function.

2. One of the solutions to address the overestimation problem in the offline framework is CQL, whose objective function for computing the value is given below. Explain the role of each of the four terms in this objective function. [20-points]

$$\begin{aligned}\hat{Q}^T = \arg\min_Q \max_\mu & \alpha \mathbb{E}_{s\sim D, a\sim\mu(a|s)}[Q(s,a)] \\ & - \alpha \mathbb{E}_{(s,a)\sim D}[Q(s,a)] \\ & - \mathbb{E}_{s\sim D}[\mathcal{H}(\mu(\cdot|s))] \\ & + \mathbb{E}_{(s,a,s')\sim D}\left[(Q(s,a) - (r(s,a) + \mathbb{E}[Q(s',a')]))^2\right]\end{aligned}$$

Answer: The function balances learning from the data with preventing overestimation for unseen actions. The roles of terms:

Term 1: This term represents the expected Q-value under a policy μ that is learned to maximize this same expectation. In practice, this means μ learns to select actions that the current Q-function believes have high values. The overall objective then minimizes the Q-function with respect to these actions, forcing the model to lower its Q-value estimates for actions that are likely out-of-distribution (OOD) but have high values.

Term2: This term represents the negated expected Q-value for state-action pairs sampled directly from the offline dataset. It maximizes the Q-values for the in-distribution actions. This prevents the Q-function from becoming overly conservative by ensuring that the values for actions known to have been taken are not excessively suppressed.

Term3: This term maximizes the entropy of the policy μ. Encouraging a high-entropy policy prevents μ from collapsing to a deterministic policy that exploits only a single high-value action. Instead, it promotes a more stochastic policy that explores the action space better, which helps identifying a wider range of potentially overestimated OOD actions.

Term4: This is the standard Mean Squared Bellman Error, a fundamental component of Q-learning. It ensures the learned Q-function is consistent with the environment's dynamics (observed in the dataset). It minimizes the difference between the current Q-value and the target Q-value derived from the reward and the expected value of the next state, grounding the entire learning process in the provided data.

3. Rewrite the optimization problem from part 3 as a minimization-only problem. [20-points]

Answer: The objective function is given as:

$$\hat{Q}^T = \arg\min_Q \max_\mu \mathbb{E}_{s\sim D, a\sim\mu(a|s)}[Q(s,a)] - \mathbb{E}_{s\sim D}[\mathcal{H}(\mu(\cdot|s))] - \alpha\mathbb{E}_{(s,a)\sim D}[Q(s,a)] + \text{MSBE} \quad (1)$$

To convert this to a minimization-only problem, the inner maximization over μ must be solved analytically.

The inner maximization problem for each state s is:

$$\max_{\mu(\cdot|s)} \left(\mathbb{E}_{a\sim\mu(a|s)}[Q(s,a)] - \mathcal{H}(\mu(\cdot|s))\right)$$

$$\max_{\mu(\cdot|s)} \left(\sum_a \mu(a|s)Q(s,a) - \sum_a \mu(a|s)\log\mu(a|s)\right) \quad \text{s.t.} \quad \sum_a \mu(a|s) = 1 \quad (2)$$

We form the Lagrangian $\mathcal{L}(\mu, \lambda)$:

$$\mathcal{L}(\mu,\lambda) = \sum_a \mu(a|s)Q(s,a) - \sum_a \mu(a|s)\log\mu(a|s) + \lambda\left(1 - \sum_a \mu(a|s)\right) \quad (3)$$

Set the partial derivative with respect to $\mu(a|s)$ to zero:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mu(a|s)} &= Q(s,a) - (\log\mu(a|s) + 1) - \lambda = 0 \\
\log\mu(a|s) &= Q(s,a) - 1 - \lambda \\
\mu(a|s) &= \exp(Q(s,a) - 1 - \lambda) \\
\mu(a|s) &= \exp(Q(s,a)) \cdot e^{-1-\lambda}
\end{aligned}$$

Using the constraint $\sum_a \mu(a|s) = 1$ to find the normalization constant:

$$\sum_a \exp(Q(s,a)) \cdot e^{-1-\lambda} = 1$$

$$e^{-1-\lambda} = \frac{1}{\sum_{a'}\exp(Q(s,a'))}$$

The optimal policy $\mu^*(a|s)$ is:

$$\mu^*(a|s) = \frac{\exp(Q(s,a))}{\sum_{a'}\exp(Q(s,a'))} \quad (4)$$

Substitute μ^* back into the maximized expression. Note that:

$$\log\mu^*(a|s) = \log\left(\frac{\exp(Q(s,a))}{\sum_{a'}\exp(Q(s,a'))}\right) = Q(s,a) - \log\sum_{a'}\exp(Q(s,a'))$$

Now compute the entropy $\mathcal{H}(\mu^*)$:

$$\begin{aligned}
\mathcal{H}(\mu^*) &= -\sum_a \mu^*(a|s) \log \mu^*(a|s) \\
&= -\sum_a \mu^*(a|s) \left(Q(s,a) - \log \sum_{a'} \exp(Q(s,a')) \right) \\
&= -\sum_a \mu^*(a|s) Q(s,a) + \left(\log \sum_{a'} \exp(Q(s,a')) \right) \left(\sum_a \mu^*(a|s) \right) \\
&= -\mathbb{E}_{a\sim\mu^*}[Q(s,a)] + \log \sum_{a'} \exp(Q(s,a'))
\end{aligned}$$

The full expression becomes:

$$\begin{aligned}
\mathbb{E}_{a\sim\mu^*}[Q(s,a)] + \mathcal{H}(\mu^*) &= \mathbb{E}_{a\sim\mu^*}[Q(s,a)] + \left(-\mathbb{E}_{a\sim\mu^*}[Q(s,a)] + \log \sum_{a'} \exp(Q(s,a')) \right) \\
&= \log \sum_{a'} \exp(Q(s,a'))
\end{aligned}$$

Replacing the inner maximization with its analytical solution, the final objective is:

$$\begin{aligned}
\hat{Q}^T = \arg\min_Q \Bigg(& \mathbb{E}_{s\sim D} \left[\log \sum_a \exp(Q(s,a)) \right] \\
& - \alpha \mathbb{E}_{(s,a)\sim D}[Q(s,a)] \\
& + \mathbb{E}_{(s,a,s')\sim D} \left[\left(Q(s,a) - \left(r(s,a) + \gamma \mathbb{E}_{a'\sim\pi(a'|s')}[Q(s',a')] \right) \right)^2 \right] \Bigg)
\end{aligned}$$

4. To apply this method in model-based reinforcement learning, what changes are needed in the objective function? Rewrite the new objective function. [10-points]

Answer: To apply CQL in a model-based setting, we first learn a dynamics model $\hat{\mathcal{T}}(s'|s,a)$ and a reward model $\hat{\mathcal{R}}(s,a)$ from the data. The primary change to the objective function is in the Mean Squared Bellman Error (MSBE) term, where the target value is computed using these learned models instead of empirical samples from the dataset.

The new model-based objective function is:

$$\begin{aligned}
\hat{Q}^T = \arg\min_Q \Bigg(& \mathbb{E}_{s\sim D} \left[\log \sum_a \exp(Q(s,a)) \right] \\
& - \alpha \mathbb{E}_{(s,a)\sim D}[Q(s,a)] \\
& + \mathbb{E}_{(s,a)\sim D} \left[\left(Q(s,a) - \left(\hat{\mathcal{R}}(s,a) + \gamma \mathbb{E}_{\substack{s'\sim\hat{\mathcal{T}}(s'|s,a) \\ a'\sim\pi(a'|s')}}[Q(s',a')] \right) \right)^2 \right] \Bigg)
\end{aligned}$$

References

[1] Cover image designed by freepik