Offline Learning for Partially Observable Markov Decision Processes

Brandon Han

May 5, 2025

1 Background

We first define a partially observable Markov decision process using the following tuple. Note that we are operating in the finite horizon setting.

$$\mathcal{G} = (\mathbb{S}, \mathbb{A}, \mathbb{O}, \mathcal{R}, \mathcal{P}, \mathcal{O}, \mu, \mathcal{H})$$

Elements $\mathbb{S}$, $\mathbb{A}$, and $\mathbb{O}$ are the set of states, actions, and observations respectively. Function $\mathcal{R} : \mathbb{S} \times \mathbb{A} \to \mathfrak{R}$ determines the reward for taking an action in a given state. $\mathcal{P} : \mathbb{S} \times \mathbb{A} \to \Delta\,\mathbb{S}$ provides the probability distribution for the next state, given the current state and the action taken. $\mathcal{O} : \mathbb{S} \to \Delta\,\mathbb{O}$ determines the probability of receiving any given observation in a specific state. μ represents the initial state distribution.

We can also use time-dependent transitions, reward functions, and emission kernels. Time-dependent quantities will be denoted by a subscript representing the timestamp.

Interaction with a partially observable Markov decision process is as follows. First, the initial state S_1 is sampled from μ. Then for each time step h until the horizon $\mathcal{H}$, the following interaction is repeated.

1. Observation O_h is sampled from $\mathcal{O}(\cdot \mid S_h)$ and is presented to the agent.
2. The agent follows some policy π to determine action A_h.
3. Reward $R_h = \mathcal{R}(S_h, A_h)$ is presented to the agent.
4. The next state S_{h+1} is sampled from $\mathcal{P}(\cdot \mid S_h, A_h)$.

We also assume that there is an initial observation O_0 sampled before S_1 is known. This quantity is conditionally independent from all other variables when given S_1.

Our goal is to find a policy with maximum value, where value is the expected cummulative reward given to the agent when operating under the given policy.

$$\mathcal{V}(\pi) = \mathbb{E}_\pi \left[\sum_{h=1}^{\mathcal{H}} R_h \mid S_1 \sim \mu \right]$$

Learning is done in an offline setting. Formally, we are given a trajectory generated by the environment and some behavioral policy π^b. Importantly, we allow the behavioral policy π^b to access the underlying state of the environment. The behavioral policy induces a probability distribution $\mathbb{P}^b$ over the space of all trajectories. We can take independent samples from $\mathbb{P}^b$ to generate trajectories.

Certain problems arise when using an offline dataset for learning. Results from causal inference state that if the elements within the dataset depend on hidden factors like S_h, then distributional shifts will occur. We unfortunately must accept results from causal inference without proof to retain our focus on reinforcement learning. For now, understand that taking empirical estimations of conditional expectations from our dataset will result in suboptimal learning.

2 Learning Setting

To handle the distributional shift, we apply the idea of confounding bridge functions from proximal causal inference to determine the value of any given policy. The first two assumptions defines the notion of the confounding bridge function and assume these functions exist.

Assumption 1 (Negative control)**.** *Under the offline data distribution, the initial observation O_0 is presampled before the decision process begins [4, 2, 3].*

$$O_0 \perp O_h, O_{h+1}, R_h \mid S_h, A_h, \Gamma_{h-1}$$

Assumption 2 (Confounding bridge functions)**.** *Let π be any history dependent policy. We assume that, for each time step h, there exists a set of value bridge functions $\{b_h^\pi : \mathbb{A} \times \mathbb{O} \times \mathbb{H} \to \mathfrak{R}\}$ that satisfy the following conditional moment equation, where $b_{\mathcal{H}+1}^\pi$ is defined to be a zero function.*

$$\begin{aligned}&\mathbb{E}_{\pi^b}\left[b_h^\pi\left(A_h, O_h, \Gamma_{h-1}\right) \mid A_h, O_0, \Gamma_{h-1}\right] \\ &= \mathbb{E}_{\pi^b}\left[\pi_h\left(A_h \mid O_h, \Gamma_{h-1}\right) \cdot \left(R_h + \sum_{a'} b_{h+1}^\pi\left(a', O_{h+1}, \Gamma_h\right)\right) \mid A_h, O_0, \Gamma_{h-1}\right]\end{aligned}$$

Assumption 3 (Completeness)**.** *For any time step h, any measurable function $g_h : \mathbb{S} \times \mathbb{A} \times \mathbb{H} \to \mathfrak{R}$ will satisfy the following with high probability.*

$$\mathbb{E}_{\pi^b}\left[g_h\left(S_h, A_h, \Gamma_h\right)\right] = 0 \qquad \Longleftrightarrow \qquad g_h\left(S_h, A_h, \Gamma_h\right) = 0$$

The completeness assumption [2, 3] ensures that some important quantities can be measured as functions of observable trajectories without knowing the underlying state of the environment. Confounding bridge functions use observable quantities to extract data from the dataset. This structure for a value bridge function represents individual components of an expectation summation and is used across all algorithms discussed here.

Theorem 1 (Value Identification)**.** *Under Assumptions 1, 2, and 3, the value associated with a policy π can be expressed as the following.*

$$\mathcal{V}\left(\pi\right) = \mathbb{E}_{\pi^b}\left[\sum_{a \in \mathbb{A}} b_1^\pi\left(a, O_1\right)\right]$$

Proof. What follows is a sketch of the proof for Theorem 1 [2]. The first step is to determine the expected values of specific desirable quantities when conditioning on the underlying state S_h. We can use the laws of conditional expectation and algebraic manipulation to rewrite the expression in Assumption 2 as the following.

$$\begin{aligned}&\mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[b_h^\pi\left(A_h, O_h, \Gamma_{h-1}\right) \mid A_h, S_h, \Gamma_{h-1}\right] \mid A_h, O_0, \Gamma_{h-1}\right] \\ &= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\pi_h\left(A_h \mid O_h, \Gamma_{h-1}\right) \cdot \left(R_h + \sum_{a'} b_{h+1}^\pi\left(a', O_{h+1}, \Gamma_h\right)\right) \mid A_h, S_h, \Gamma_{h-1}\right] \mid A_h, O_0, \Gamma_{h-1}\right]\end{aligned}$$

The inner conditional expectations are functions of A_h, S_h, and Γ_{h-1}, so they are members of the same function class as the statement of Assumption 3. As a result, we can use the assumption to state the following. Note that although similar to Assumption 2, this statement conditions on the current state of the

environment, rather than the initial observation.

$$\begin{aligned}
&\mathbb{E}_{\pi^b}\left[b_h^\pi\left(A_h, O_h, \Gamma_{h-1}\right) \mid A_h, S_h, \Gamma_{h-1}\right] \\
&= \mathbb{E}_{\pi^b}\left[\pi_h\left(A_h \mid O_h, \Gamma_{h-1}\right) \cdot \left(R_h + \sum_{a'} b_{h+1}^\pi\left(a', O_{h+1}, \Gamma_h\right)\right) \mid A_h, S_h, \Gamma_{h-1}\right]
\end{aligned}$$

Next, we can use this result to state the following as a lemma. This statement can be proved by applying induction on the time step h using $h = \mathcal{H}$ as the base case.

$$\mathbb{E}_{\pi^b}\left[\sum_{a \in \mathbb{A}} b_h^\pi\left(a, O_h, \Gamma_{h_1}\right) \mid S_h, \Gamma_{h-1}\right] = \mathbb{E}_\pi\left[\sum_{j=h}^{\mathcal{H}} R_j \mid S_h, \Gamma_{h-1}\right]$$

After proving this lemma, the correctness of our theorem follows easily. Note that Γ_0 is an empty set.

$$\begin{aligned}
\mathcal{V}^\pi &= \mathbb{E}_\pi\left[\sum_{h=1}^{\mathcal{H}} R_h\right] = \mathbb{E}_{S_1 \sim \mu}\left[\mathbb{E}_\pi\left[\sum_{h=1}^{\mathcal{H}} R_h \mid S_1, \Gamma_0\right]\right] \\
&= \mathbb{E}_{S_1 \sim \mu}\left[\mathbb{E}_{\pi^b}\left[\sum_{a \in \mathbb{A}} b_1\left(a, O_1, \Gamma_0\right) \mid S_1, \Gamma_0\right]\right] = \mathbb{E}_{\pi^b}\left[\sum_{a \in \mathbb{A}} b_1\left(a, O_1\right)\right]
\end{aligned}$$

□

This theorem is important because it states that if we have the value bridge function associated with a policy, we can calculate the expected value of the policy.

2.1 Estimating the Value Bridge Function

It then remains to calculate the value bridge function. To do this, our algorithms use a minimax estimation procedure introduced by Dikkala et al [1]. More specifically, we can use backward induction on the time step h and calculate $\hat{b}_h^\pi$ when given b_{h+1}^π by minimizing a loss function $\mathcal{L}_h^\pi$. First, we can define the following.

$$\begin{aligned}
&l_h^\pi\left(\hat{b}_h^\pi, b_{h+1}^\pi\right)\left(A_h, O_0, \Gamma_{h-1}\right) \\
&= \mathbb{E}_{\pi^b}\left[\hat{b}_h^\pi\left(A_h, O_h, \Gamma_{h-1}\right) - \pi_h\left(A_h \mid O_h, \Gamma_{h-1}\right) \cdot \left(R_h + \sum_{a'} b_{h+1}^\pi\left(a', O_{h+1}, \Gamma_h\right)\right) \mid A_h, O_0, \Gamma_{h-1}\right]
\end{aligned}$$

Then we can define our loss function $\mathcal{L}_h^\pi$ to be the expected squared magnitude of this difference, where the expectation is taken over A_h, O_0, and Γ_{h-1}.

$$\mathcal{L}_h^\pi\left(\hat{b}_h^\pi, b_{h+1}^\pi\right) = \mathbb{E}_{\pi^b}\left(l_h^\pi\left(\hat{b}_h^\pi, b_{h+1}^\pi\right)\left(A_h, O_0, \Gamma_{h-1}\right)\right)^2$$

However, we still face the issue that using the empirical estimates from the dataset to minimize this loss will result in distributional shifts. To resolve this issue, we apply the Fenchel duality and the interchangeability principle to rewrite the loss function as the following, where $\lambda > 0$ and $\mathbb{G}$ is a function class to be discussed below.

$$\begin{aligned}
\mathcal{L}_h^\pi\left(\hat{b}_h^\pi, b_{h+1}^\pi\right) &= 4\lambda \cdot \max_{g \in \mathbb{G}} \Phi_{\pi,h}^\lambda\left(\hat{b}_h^\pi, b_{h+1}^\pi, g\right) \\
\Phi_{\pi,h}^\lambda\left(\hat{b}_h^\pi, b_{h+1}, g\right) &= \mathbb{E}_{\pi^b}\left[l_h^\pi\left(\hat{b}_h^\pi, b_{h+1}^\pi\right)\left(A_h, O_0, \Gamma_{h-1}\right) \cdot g\left(A_h, O_0, \Gamma_{h-1}\right) - \lambda g\left(A_h, O_0, \Gamma_{h-1}\right)^2\right]
\end{aligned}$$

This is stated without proof because it strays too far from our focal topic of reinforcement learning. However, it is worth having an intuitive understanding of this claim. We introduce an adversary to our learner that

would like to highlight the A_h, O_0, Γ_{h-1} in which our learner is performing poorly. The adversary maximizes a weighted sum of $l_h^\pi\left(b_h^\pi, b_{h+1}^\pi\right)$ across every possible trajectory A_h, O_0, Γ_{h-1}, scaled by the probability that the trajectory occurs. The adversary must also work under a budget, as determined by the parameter λ.

The results from Dikkala et al [1] state that minimizing the empirical estimate of this revised loss function allows for a fast statistical rate of convergence as dictated by $\tilde{O}\left(n^{-1/2}\right)$. Then we can define the following operator to empirically estimate b_h^π when given b_{h+1}^π.

$$\hat{b}_h^\pi\left(b_{h+1}^\pi\right) \in \underset{b \in \mathbb{B}}{\operatorname{argmin}} \left[\max_{g \in \mathbb{G}} \hat{\Phi}_{\pi,h}^\lambda\left(b, b_{h+1}^\pi, g\right)\right]$$

The algorithms we will discuss use this method of minimax estimation to determine the value bridge function for a given policy at every time step, with some mild technical assumptions about the function classes $\mathbb{B}$ and $\mathbb{G}$ illustrated below.

Assumption 4 (Function classes)**.** *We assume the following about the function classes $\mathbb{B}$ and $\mathbb{G}$.*

- *All functions in $\mathbb{B}$ are bounded above by $M_{\mathbb{B}}$ and all functions in $\mathbb{G}$ are bounded above by $M_{\mathbb{G}}$.*
- *$\mathbb{G}$ is star-shaped and symmetric, i.e. $cg \in \mathbb{G}$ for all $g \in \mathbb{G}$ and $c \in [-1, 1]$.*
- *The localized population Rademacher complexity of $\mathbb{G}$ with radius α is bounded above by $\frac{\alpha^2}{M_{\mathbb{G}}}$.*
- *The function class $\mathbb{G}$ is complete and the function class $\mathbb{B}$ is realizable, as defined below.*

$$\frac{1}{2\lambda} l_h^\pi\left(b, b_{h+1}\right) \in \mathbb{G} \qquad b_h^\pi \in \mathbb{B} \qquad \forall h \in [\mathcal{H}] \quad \forall b_h, b_{h+1} \in \mathbb{B} \quad \forall \pi \in \Pi$$

3 Pessimism

The first algorithm we will discuss creates a confidence region around the minimax estimation and applies pessimism to choose a policy. First, define $\mathbf{b}^\pi = \left(b_1^\pi, \cdots, b_{\mathcal{H}}^\pi\right)$. For each policy π, construct the confidence region as follows.

$$\mathrm{CR}^\pi(\epsilon) = \left\{\mathbf{b} \in \mathbb{B} \times \cdots \times \mathbb{B} \mid \max_{h \in [\mathcal{H}]} \left[\max_{g \in \mathbb{G}} \hat{\Phi}_{\pi,h}^\lambda\left(b_h, b_{h+1}^\pi, g\right) - \max_{g \in \mathbb{G}} \hat{\Phi}_{\pi,h}^\lambda\left(b_h^\pi, b_{h+1}^\pi, g\right)\right] \leq \epsilon\right\}$$

The choice of ϵ determines the size of the confidence region. We then determine the pessimistic estimate of the policy's value by taking the worst possible value bridge function that lies within the confidence region. The output of this pessimism algorithm is the policy that maximizes this pessimistic estimate.

$$\hat{\mathcal{V}}(\pi) = \min_{\mathbf{b} \in \mathrm{CR}^\pi(\epsilon)} \mathbb{E}_{\pi^b}\left[\sum_{a \in \mathbb{A}} b_1(a, O_1)\right] \qquad \hat{\pi} \in \underset{\pi \in \Pi}{\operatorname{argmax}}\, \hat{\mathcal{V}}(\pi)$$

With the standard partial coverage assumption, we now seek to prove the optimality of this choice.

Assumption 5 (Partial coverage)**.** *We assume that for any policy π and any time step h, there exists a set of confounding bridge functions $\{q_h^\pi : \mathbb{A} \times \mathbb{O} \to \mathfrak{R}\}$ that satisfy the following.*

$$\mathbb{E}_{\pi^b}\left[q_h^\pi(A_h, O_0) \mid A_h, S_h, \Gamma_{h-1}\right] = \frac{\mu_h(S_h, \Gamma_{h-1})}{\pi_h^b(A_h \mid S_h)} \qquad \mu_h(S_h, \Gamma_{h-1}) = \frac{\mathbb{P}_h^\pi(S_h, \Gamma_{h-1})}{\mathbb{P}_h^{\pi^b}(S_h, \Gamma_{h-1})}$$

Additionally, we define the concentrability coefficient C^{π^*} for the optimal policy π^* as the following and assume that C^{π^*} is finite.

$$C^{\pi^*} = \max_{h \in [\mathcal{H}]} \mathbb{E}_{\pi^b}\left(q_h^{\pi^*}(A_h, O_0)\right)^2 < \infty$$

Theorem 2 (Optimality of pessimism)**.** *Choose ϵ as follows, where C_1, C_1', c_1, and c_2 are global constants.*

$$\epsilon = \frac{1}{n} \cdot C_1 \cdot M_{\mathbb{B}}^2 M_{\mathbb{G}}^2 \cdot \psi \qquad\qquad \psi = \log \frac{|\mathbb{B}| \cdot |\Pi| \cdot \mathcal{H}}{\min\left\{\delta, 4c_1 \cdot e^{-c_2 n \alpha^2}\right\}}$$

Under Assumptions 1, 2, 3, 4, and 5, we have the following guarantee.

$$SubOpt(\hat{\pi}) \leq C_1' \, \mathcal{H} \, M_{\mathbb{B}} M_{\mathbb{G}} \sqrt{\frac{C^{\pi^*} \psi}{n}}$$

Proof. What follows is a brief sketch of Theorem 2, as shown by Lu et at [4].

First, for notational convenience, define the following, where $\hat{\mathbb{E}}$ indicates an empirical measurement.

$$F(\mathbf{b}) = \mathbb{E}_{\pi^b}\left[\sum_{a \in \mathbb{A}} b_1^{\pi}(a, O_1)\right] \qquad\qquad \hat{F}(\mathbf{b}) = \hat{\mathbb{E}}_{\pi^b}\left[\sum_{a \in \mathbb{A}} b_1^{\pi}(a, O_1)\right]$$

The estimation for the value bridge requires a backpropagation through time. We must show that the error from the first estimation does not explode as we backpropagate. This is summarized through the following lemma, which states that the error reflected in the value measurement is bounded by the sum of the local error at each time step.

Lemma 1. *Take any arbitrary π and consider its true value bridge function $\mathbf{b}^{\pi}$. Let C^{π} be the concentrability coefficient for π under the behavioral policy. Now consider any value bridge function $\mathbf{b}$. We can show the following.*

$$F(\mathbf{b}^{\pi}) - F(\mathbf{b}) \leq \sum_{h=1}^{\mathcal{H}} \sqrt{C^{\pi} \cdot \mathcal{L}_h^{\pi}(b_h, b_{h+1})}$$

We also must determine the validity of the established confidence regions, which can be shown as the following.

Lemma 2. *We choose the radius ϵ of our confidence regions as the following, where ψ is defined in Theorem 2.*

$$\epsilon = C_1 \frac{\lambda + 1}{\lambda} \cdot M_{\mathbb{B}}^2 M_{\mathbb{G}}^2 \cdot \frac{\psi}{n}$$

Then with probability greater than $1 - \delta$, the true value bridge function $\mathbf{b}^{\pi}$ lies within the confidence interval. Furthermore, we can bound the error of the minimax estimation at each time step by the following, where $\tilde{C}$ is a global constant.

$$\sqrt{\mathcal{L}_h^{\pi}(b_h, b_{h+1})} \leq \tilde{C}_1 M_{\mathbb{B}} M_{\mathbb{G}} \sqrt{\frac{\lambda + 1}{\lambda} \cdot \frac{\psi}{n}}$$

With these results, we can prove Theorem 2. First, decompose the suboptimality into three separate components.

$$\begin{aligned} \text{SubOpt}(\hat{\pi}) &= \mathcal{V}^{\pi^*} - \mathcal{V}^{\hat{\pi}} = F\left(\mathbf{b}^{\pi^*}\right) - F\left(\mathbf{b}^{\hat{\pi}}\right) \\ &\leq \left[F\left(\mathbf{b}^{\pi^*}\right) - \hat{F}\left(\mathbf{b}^{\pi^*}\right)\right] + \left[F\left(\mathbf{b}^{\pi^*}\right) - \hat{F}\left(\mathbf{b}^{\hat{\pi}}\right)\right] + \left[\hat{F}\left(\mathbf{b}^{\hat{\pi}}\right) - F\left(\mathbf{b}^{\hat{\pi}}\right)\right] \end{aligned}$$

The outer terms can be bounded by the Hoeffding bound and concentration inequalities. To bound the inner term, we condition upon the event that the true value bridge function lies within the confidence region and apply the definition of $\hat{\pi}$. Using these substitutions, we gain the following.

$$\begin{aligned}
\text{SubOpt}\left(\hat{\pi}\right) &\leq \left[\max_{\mathbf{b}\in \text{CR}^{\pi^*}(\epsilon)} \hat{F}\left(\mathbf{b}\right) - \max_{\pi\in\Pi(\epsilon)} \min_{\mathbf{b}\in \text{CR}^{\pi}(\epsilon)} \hat{F}\left(\mathbf{b}\right)\right] + 2\sqrt{\frac{2M_{\mathbb{B}}^2}{n}\log\frac{|\mathbb{B}|}{\delta}} \\
&\leq \left[\max_{\mathbf{b}\in \text{CR}^{\pi^*}(\epsilon)} \hat{F}\left(\mathbf{b}\right) - \min_{\mathbf{b}\in \text{CR}^{\pi^*}(\epsilon)} \hat{F}\left(\mathbf{b}\right)\right] + 2\sqrt{\frac{2M_{\mathbb{B}}^2}{n}\log\frac{|\mathbb{B}|}{\delta}} \\
&\leq 2\max_{\mathbf{b}\in \text{CR}^{\pi^*}(\epsilon)} \left|\hat{F}\left(\mathbf{b}\right) - \hat{F}\left(\mathbf{b}^{\pi^*}\right)\right| + 2\sqrt{\frac{2M_{\mathbb{B}}^2}{n}\log\frac{|\mathbb{B}|}{\delta}}
\end{aligned}$$

The next step is to apply the triangle inequality to split the remaining complicated term into three separate components. the two outer components can again be bounded by concentration inequalities.

$$\begin{aligned}
&\text{SubOpt}\left(\hat{\pi}\right) \\
&\leq 2\max_{\mathbf{b}\in \text{CR}^{\pi^*}(\epsilon)} \left|\hat{F}\left(\mathbf{b}\right) - \hat{F}\left(\mathbf{b}^{\pi^*}\right)\right| + 2\sqrt{\frac{2M_{\mathbb{B}}^2}{n}\log\frac{|\mathbb{B}|}{\delta}} \\
&\leq 2\max_{\mathbf{b}\in \text{CR}^{\pi^*}(\epsilon)} \left|\hat{F}\left(\mathbf{b}\right) - F\left(\mathbf{b}\right)\right| + 2\max_{\mathbf{b}\in \text{CR}^{\pi^*}(\epsilon)} \left|F\left(\mathbf{b}\right) - F\left(\mathbf{b}^{\pi^*}\right)\right| + 2\left|F\left(\mathbf{b}^{\pi^*}\right) - \hat{F}\left(\mathbf{b}^{\pi^*}\right)\right| + 2\sqrt{\frac{2M_{\mathbb{B}}^2}{n}\log\frac{|\mathbb{B}|}{\delta}} \\
&\leq 2\max_{\mathbf{b}\in \text{CR}^{\pi^*}(\epsilon)} \left|F\left(\mathbf{b}\right) - F\left(\mathbf{b}^{\pi^*}\right)\right| + 4\sqrt{\frac{2M_{\mathbb{B}}^2}{n}\log\frac{|\mathbb{B}|}{\delta}}
\end{aligned}$$

Finally, we apply the two lemmas to complete the proof, using the constant C_1' to absorb multiplicative factors.

$$\begin{aligned}
\text{SubOpt}\left(\hat{\pi}\right) &\leq 2\max_{\mathbf{b}\in \text{CR}^{\pi^*}(\epsilon)} \left|F\left(\mathbf{b}\right) - F\left(\mathbf{b}^{\pi^*}\right)\right| + 4\sqrt{\frac{2M_{\mathbb{B}}^2}{n}\log\frac{|\mathbb{B}|}{\delta}} \\
&\leq 2\sum_{h=1}^{\mathcal{H}} \sqrt{C^{\pi^*}\cdot\mathcal{L}_h^{\pi^*}\left(b_h, b_{h+1}\right)} + 4\sqrt{\frac{2M_{\mathbb{B}}^2}{n}\log\frac{|\mathbb{B}|}{\delta}} \\
&\leq 2\sum_{h=1}^{\mathcal{H}} \sqrt{C^{\pi^*}}\cdot\left(\tilde{C}_1 M_{\mathbb{B}} M_{\mathbb{G}}\sqrt{\frac{\lambda+1}{\lambda}\cdot\frac{\psi}{n}}\right) + 4\sqrt{\frac{2M_{\mathbb{B}}^2}{n}\log\frac{|\mathbb{B}|}{\delta}} \\
&= 2\,\mathcal{H}\sqrt{C^{\pi^*}}\cdot\left(C_1' M_{\mathbb{B}} M_{\mathbb{G}}\sqrt{\frac{\psi}{n}}\right) + 4\sqrt{\frac{2M_{\mathbb{B}}^2}{n}\log\frac{|\mathbb{B}|}{\delta}} \\
&\leq C_1'\,\mathcal{H}\,M_{\mathbb{B}} M_{\mathbb{G}}\sqrt{\frac{C^{\pi^*}\psi}{n}}
\end{aligned}$$

The last inequality holds because the second term takes a similar form as ψ, so the two terms can be combined, resulting in a hidden multiplicative factor absorbed by C_1'.

□

With this, we reconsider ψ. Notice that ψ actually increases polynomially with n due to the exponential factor in the denominator. This would imply that the suboptimality does not decrease with n. However, Lu et at [4] shows that we can simply reduce parameter α to counteract this.

$$\psi = \log\frac{|\mathbb{B}|\cdot|\Pi|\cdot\mathcal{H}}{\min\left\{\delta, 4c_1\cdot e^{-c_2 n\alpha^2}\right\}} \to c_2 n\alpha^2\log\frac{|\mathbb{B}|\cdot|\Pi|\cdot\mathcal{H}}{\min\left\{\delta, 4c_1\right\}} \qquad\qquad \alpha = O\left(\frac{1}{\sqrt{n}}\right)$$

4 Improvements

To further develop this field, Hong et al [2, 3] introduced new ways to apply the minimax estimation procedure. The goal of their improvements was to create a more computationally feasible method, addressing the intractibility of certain minimizations used by Lu et al [4].

4.1 Policy Gradient

Namely, we decide to identify new classes of confounding bridge functions to capture different quantities from the dataset. Recall that the value bridge function was used to identify the value of a policy. Instead, we can identify the policy gradient. For any π_θ, we assume that there exists some function $f_h^{\pi_\theta}$ that satisfies the following.

$$\begin{aligned}
&\mathbb{E}_{\pi^b}\left[f_h^{\pi_\theta}\left(A_h, O_h, \Gamma_{h-1}\right) \mid A_h, \Gamma_{h-1}, 0_0\right] \\
&= \mathbb{E}_{\pi^b}\left[\left(R_h + \sum_{a'} b_{h+1}^{\pi_\theta}\left(a', O_{h+1}, \Gamma_h\right)\right) \nabla_\theta \pi_\theta\left(A_h \mid O_h, \Gamma_{h-1}\right) + \sum_{a'} f_{h+1}^{\pi_\theta}\left(a', O_{h+1}, \Gamma_h\right) \pi_\theta\left(A_h \mid O_h, \Gamma_{h-1}\right) \mid A_h, \Gamma_{h-1}, O_0\right]
\end{aligned}$$

This gradient bridge function takes the form of the derivative of the value bridge function after applying the product rule for differentiation. It can then be used to determine the gradient of the value with respect to the policy's parameters.

$$\nabla_\theta \mathcal{V}\left(\pi_\theta\right) = \mathbb{E}_{\pi^b}\left[\sum_{a \in \mathbb{A}} f_1^{\pi_\theta}\left(a, O_1\right)\right]$$

The optimality of using this gradient estimation requires additional assumptions. If the behavioral policy has full coverage and the parameter is convex and sufficiently smooth, we can establish a suboptimal guarantee for the gradient descent algorithm [2].

$$\mathcal{V}^{\pi^*} - \max_k \mathcal{V}^{\pi_{\theta_k}} = O\left(\frac{1}{\sqrt{K}} + \frac{1}{\sqrt{n}}\right) + \epsilon_{approx}$$

As with the online version of policy gradient, this guarantee only applies for one policy in the gradient descent trajectory, without any statement on which.

References

[1] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models, 2020.

[2] Mao Hong, Zhengling Qi, and Yanxun Xu. A policy gradient method for confounded pomdps, 2023.

[3] Mao Hong, Zhengling Qi, and Yanxun Xu. Model-based reinforcement learning for confounded POMDPs. In *Forty-first International Conference on Machine Learning*, 2024.

[4] Miao Lu, Yifei Min, Zhaoran Wang, and Zhuoran Yang. Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes, 2024.