Confounding-Robust Fitted-Q-Iteration under Observed Markovian Marginals

David Bruns-Smith
Stanford University
causal@stanford.edu

Angela Zhou
University of Southern California
zhoua@usc.edu

Abstract

Offline reinforcement learning (RL) methods like fitted-Q-evaluation/iteration (FQE/FQI) are popular for learning policies from observational data, but they assume all decision-relevant covariates are observed. In practice, unobserved confounders can bias estimates and lead to harmful policies. We develop a robust extension of FQE/FQI that provides bounds on policy values under unobserved confounding via sensitivity analysis. Our key insight is that when observed state transitions are Markovian, the decision problem remains a tractable MDP even with unobserved confounders, enabling robust optimization. We introduce an orthogonalized estimation procedure that significantly improves performance over naive approaches. Experiments on healthcare data demonstrate the method's effectiveness in learning robust policies from confounded observational data.

1 Introduction and Problem Setup

Sequential decision-making in medicine, economics, and e-commerce often requires learning from historical observational data when online experimentation is costly or unethical. Offline reinforcement learning methods, particularly fitted-Q-evaluation and fitted-Q-iteration (FQE/FQI), have gained popularity for their computational efficiency and scalability [6]. However, these methods assume all decision-relevant covariates are observed. In practice, unobserved confounders—factors that influence both historical decisions and outcomes—introduce bias and can lead to harmful policies. For example, physicians may base treatment decisions on unrecorded factors like patient affect or clinical intuition, not just lab values. Our goal is to learn optimal sequential decision policies robust to unobserved confounding. We build on sensitivity analysis techniques from causal inference, which parameterize confounding strength via its effect on treatment selection probabilities [2, 3]. We adopt the marginal sensitivity model (MSM) of Tan, which has been widely used for single-timestep policy optimization. Our key contributions are as follows: (1) We show that when observed state transitions are Markovian, the decision problem remains a tractable MDP even with unobserved confounders, avoiding the computational challenges of POMDPs. (2) We develop robust FQE/FQI that provides bounds on policy values under confounding. (3) We introduce orthogonalized estimation that dramatically improves performance over naive approaches. (4) We demonstrate the method on real healthcare data for sepsis management. (5) We introduce warmstarting based on confounded bounds.

We consider a finite-horizon MDP $\mathcal{M} = (\mathcal{S} \times \mathcal{U}, \mathcal{A}, r, P, \chi, T)$ where $\mathcal{S}$ are observed states, $\mathcal{U}$ are unobserved confounders, $\mathcal{A}$ is the action space, r are rewards, P are transition probabilities, χ is the initial state distribution, and T is the horizon.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

A policy π maps states and confounders to action distributions: $\pi_t(a|s,u) = P(A_t = a|S_t = s, U_t = u)$. The value function under policy π is:

$$V_t^\pi(s,u) = \mathbb{E}\left[\sum_{k=t}^{T-1} r_k(S_k, A_k, S_{k+1}) \mid S_t = s, U_t = u, \pi\right] \tag{1}$$

We want to find a policy π^e that depends only on observed states and maximizes the marginal value:

$$V_0^{\pi^e} = \mathbb{E}_{S_0,U_0\sim\chi}[V_0^{\pi^e}(S_0, U_0)] \tag{2}$$

Our analysis requires the following assumptions:

Assumption 1 (Observed-State Markov Property)**.** *Let $H_t \coloneqq (S_{t-1}, A_{t-1}, ..., S_0, A_0)$ be the history of observed variables before time t. Then for all s, a, h, t:*

$$P_{obs}(S_{t+1}|S_t = s, A_t = a, H_t = h) = P_{obs}(S_{t+1}|S_t = s, A_t = a),$$
$$P_{obs}(A_t|S_t = s, H_t = h) = P_{obs}(A_t|S_t = s).$$

Assumption 2 (Faithfulness, Informal)**.** *Two random variables X_1 and X_2 in the underlying MDP are conditionally independent given a set of variables X_S if and only if there are no unblocked backdoor paths from X_1 to X_2 given X_S.*

These assumptions together ensure that the online decision problem remains a tractable MDP even with unobserved confounders.

Proposition 1 (Marginal MDP)**.** *Let $\chi^m \in \Delta(\mathcal{S})$ be the marginal distribution of S_0 under χ. Given Assumptions 1 and 2, there exists transition probabilities $P_t^m : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ on the observed state such that for any π^e that does not depend on U_t, the full information MDP running the policy π^e is equivalent to the* marginal MDP*, $(\mathcal{S}, \mathcal{A}, r, P^m, \chi^m, T)$, running the policy π^e.*

Memoryless Confounders An important special case is memoryless unobserved confounders:

Definition 1 (Memoryless Unobserved Confounders)**.** *The full-information MDP has* memoryless unobserved confounders *if U_t is independent of $S_{t-1}, U_{t-1}, A_{t-1}$ given S_t.*

Proposition 2.
1. *Assuming memoryless confounding alone is sufficient for all results in this paper.*
2. *Under non-trivial confounding, Assumptions 1 and 2 imply memoryless confounding.*

Please see the full paper for higher-order generalizations.

2 Method

2.1 Bias Characterization from Unobserved Confounding

We first characterize the exact bias from confounding for estimating conditional expectations in the online marginal MDP.

Theorem 1 (Confounding for Regression)**.** *Define the marginal behavior policy, $\pi_t^b(a|s) \coloneqq P_{obs}(A_t = a|S_t = s)$. Let $f(s, a, s')$ be any function. Given Assumptions 1 and 2, for all s, a, t,*

$$\mathbb{E}_{P_t^m}\left[f(S_t, A_t, S_{t+1})|S_t = s, A_t = a\right] = \mathbb{E}_{obs}\left[\frac{\pi_t^b(A_t|S_t)}{\pi_t^b(A_t|S_t, U_t)} f(S_t, A_t, S_{t+1}) \mid S_t = s, A_t = a\right].$$

The bias is potentially unbounded since U_t is unobserved. We next derive robust MDPs from restrictions on the strength of unobserved confounding.

2.2 Robust Q-Functions via Sensitivity Analysis

We adopt the marginal sensitivity model:

Assumption 3 (Marginal Sensitivity Model)**.** *There exists Λ such that $\forall t, s \in \mathcal{S}, u \in \mathcal{U}, a \in \mathcal{A}$,*

$$\Lambda^{-1} \leq \left(\frac{\pi_t^b(a \mid s, u)}{1 - \pi_t^b(a \mid s, u)}\right) / \left(\frac{\pi_t^b(a \mid s)}{1 - \pi_t^b(a \mid s)}\right) \leq \Lambda. \tag{3}$$

Algorithm 1 Confounding-Robust Fitted-Q-Iteration

1: Estimate the marginal behavior policy $\pi_t^b(a|s)$. Compute $\{\alpha_t(S_t^{(i)}, A_t^{(i)})\}_{i=1}^n$ as in **??**. Initialize $\hat{}_T = 0$.
2: **for** $t = T-1, \ldots, 1$ **do**
3: Compute the nominal outcomes $\{Y_t^{(i)}(\hat{}_{t+1})\}_{i=1}^n$ as in **??**.
4: For $a \in \mathcal{A}$, where $A_t^{(i)} = a$, fit $\hat{}_t^{1-\tau}$ the $(1-\tau)$th conditional quantile of the outcomes $Y_t^{(i)}$.
5: Compute pseudooutcomes $\{\tilde{Y}_t^{(i)}(\hat{}_t^{1-\tau}, \hat{}_{t+1})\}_{i=1}^n$ as in **??**.
6: For $a \in \mathcal{A}$, where $A_t^{(i)} = a$, fit $\hat{}_t$ via least-squares regression of $\hat{}_t^{(i)}$ against $(S_t^{(i)}, A_t^{(i)})$.
7: Compute $\pi_t^*(s) \in \arg\max_a \hat{}_t(s,a)$.
8: **end for**

This implies bounds on the unobserved ratio:

$$\alpha_t(S_t, A_t) \leq \frac{\pi_t^b(A_t|S_t)}{\pi_t^b(A_t|S_t, U_t)} \leq \beta_t(S_t, A_t) \tag{4}$$

where $\alpha_t(S_t, A_t) \coloneqq \pi_t^b(A_t|S_t) + \Lambda^{-1}(1 - \pi_t^b(A_t|S_t))$ and $\beta_t(S_t, A_t) \coloneqq \pi_t^b(A_t|S_t) + \Lambda(1 - \pi_t^b(A_t|S_t))$.

Proposition 3. *Define the uncertainty set:*

$$\mathcal{P}_t^{s,a} \coloneqq \left\{ \bar{P}_t(\cdot \mid s,a) \colon \alpha_t(s,a) \leq \frac{\bar{P}(s_{t+1} \mid s,a)}{P_{obs}(s_{t+1} \mid s,a)} \leq \beta_t(s,a), \forall s_{t+1}; \int \bar{P}_t(s_{t+1} \mid s,a) ds_{t+1} = 1 \right\}$$

Under Assumptions 1, 2, and 3, we have that $P_t^m \in \mathcal{P}_t$.

Robust Bellman Operators

Definition 2 (Robust Bellman Operators)**.** *For any function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,*

$$(\bar{\mathcal{T}}_t^{\pi^e} Q)(s,a) \coloneqq \inf_{\bar{P}_t \in \mathcal{P}_t} \mathbb{E}_{\bar{P}_t}[R_t + Q(S_{t+1}, \pi_{t+1}^e) | S_t = s, A_t = a], \tag{5}$$

$$(\bar{\mathcal{T}}_t^{*} Q)(s,a) \coloneqq \inf_{\bar{P}_t \in \mathcal{P}_t} \mathbb{E}_{\bar{P}_t}[R_t + \max_{A'}\{Q(S_{t+1}, A')\} | S_t = s, A_t = a]. \tag{6}$$

Proposition 4 (Robust Bellman Equation)**.** *Let $|\mathcal{A}| = 2$ and let Assumptions 1, 2, and 3 hold. Then for any π, $\bar{Q}_t^\pi(s,a) = \bar{\mathcal{T}}_t^\pi \bar{Q}_{t+1}^\pi(s,a)$ and $\bar{V}_t^\pi(s) = \mathbb{E}_{A \sim \pi_t(s)}[\bar{Q}_t^\pi(s,A)]$.*

The robust Bellman operator has a closed-form solution involving conditional quantiles:

Proposition 5 (Conditional expected shortfall closed form solution)**.** *Define $\tau \coloneqq \Lambda/(1+\Lambda)$ and $Y_t(Q) \coloneqq R_t + \max_{a'}[Q(S_{t+1}, a')]$. The robust Bellman operator admits the closed-form solution:*

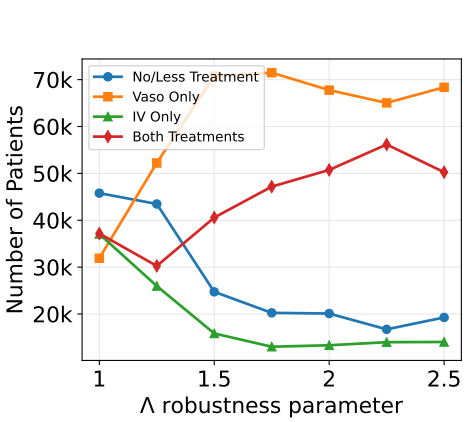
$$(\bar{\mathcal{T}}_t^{*} Q)(s,a) = \mathbb{E}_{obs}\left[\alpha_t Y_t(Q) + \frac{1-\alpha_t}{1-\tau} Y_t(Q) \mathbb{I}\{Y_t(Q) \leq Z_{t,a}^{1-\tau}\} \mid S_t = s, A_t = a\right] \tag{7}$$

where $Z_{t,a}^{1-\tau}$ is the $(1-\tau)$-level conditional quantile of $Y_t(Q)$ given $S_t = s, A_t = a$.

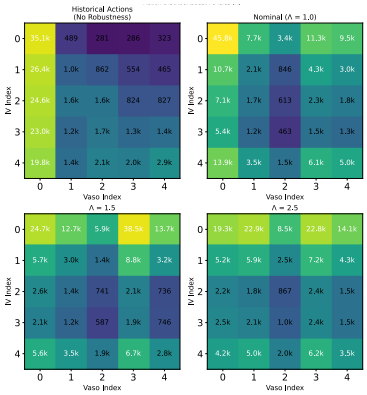
Orthogonalized Estimation The two-stage procedure depends on the conditional quantile function, a nuisance function that must be estimated. To avoid transferring biased first-stage estimation error, we introduce orthogonalization:

$$\tilde{Y}_t(Z, Q) \coloneqq \alpha_t Y_t(Q) + \frac{1-\alpha_t}{1-\tau}\Big(Y_t(Q)\mathbb{I}\{Y_t(Q) \leq Z\} - Z \cdot \{\mathbb{I}\{Y_t(Q) \leq Z\} - (1-\tau)\}\Big) \tag{8}$$

This orthogonalized pseudo-outcome is Neyman-orthogonal with respect to error in the quantile function, enabling $\sqrt{n}$-consistent estimation.



(a) Counts of actions taken by the robust optimal policy vs. the sensitivity parameter Λ.



(b) Heatmaps of $\log(\text{action counts}) + 1$. Yellow is higher, blue is lower count. IV action index is on the y-axis, vasopressor is on the x-axis; dosage increases in action index.

Figure 1: Summary and heatmaps of optimal actions as Λ increases.

Analysis Please see the paper for full theorem statements and analysis. Orthogonality permits the following improvement in estimation; the error in estimating quantiles is second order for estimation of Q.

Proposition 6 (CVaR estimation error)**.** . *Assume Assumptions 4 to 6. For $a \in \mathcal{A}, t \in [T-1]$, if the conditional quantile estimation is $o_p\left(n^{-\frac{1}{4}}\right)$ consistent, i.e. $\left\|\hat{Z}_t^{1-\tau} - Z_t^{1-\tau}\right\|_\infty = o_p\left(n^{-\frac{1}{4}}\right), \mathbb{E}\left[\|\hat{Z}_t^{1-\tau} - Z_t^{1-\tau}\|_2\right] = o_p\left(n^{-\frac{1}{4}}\right)$, then*

$$\left\|\hat{\bar{Q}}_t(S,a) - \bar{Q}_t(S,a)\right\|_2 \leq \left\|\widetilde{\bar{Q}}_t(S,a) - \bar{Q}_t(S,a)\right\|_2 + o_p\left(n^{-\frac{1}{2}}\right).$$

The results depend on the functional complexity of Q and quantile function spaces.

Assumption 4 (Finite function classes.)**.** *The Q-function class $\mathcal{Q}$ and conditional quantile class $\mathcal{Z}$ are finite but can be exponentially large.*

Assumption 5 (Infinite function classes with well-behaved covering number.)**.** *The Q-function class $\mathcal{Q}$, and conditional quantile class $\mathcal{Z}$ have covering numbers $N(\epsilon, \mathcal{Q}, d)$, $N(\epsilon, \mathcal{Z}, d)$ (respectively).*

Theorem 2 (Fitted Q Iteration guarantee)**.** *Suppose there is a continuous density, concentrability, regression stability, product-rate error rates, Bellman completeness, and let B_R be a bound on rewards. Recall that $\mathcal{E}(\hat{Q}) = \frac{1}{T}\sum_{t=0}^{T-1}\left\|\hat{Q}_t -_t^\star \hat{Q}_{t+1}\right\|_{\mu_t}^2$. Then, with probability greater than $1-\delta$, under Assumption 4 (finite function class), we have that*

$$\mathcal{E}(\hat{Q}) \leq \epsilon_{\mathcal{Q},\mathcal{Z}} + \frac{56(T^2+1)B_R \log\{T|\mathcal{Q}||\mathcal{Z}|/\delta\}}{3n} + \sqrt{\frac{32(T^2+1)B_R \log\{T|\mathcal{Q}||\mathcal{Z}|/\delta\}}{n}\epsilon_{\mathcal{Q},\mathcal{Z}}} + o_p(n^{-1}),$$

while under Assumption 5 (infinite function class), choosing the covering number approximation error $\epsilon = O(n^{-1})$ such that $\epsilon_{\mathcal{Q},\mathcal{Z}} = O(n^{-1})$, we have that

$$\mathcal{E}(\hat{Q}) \leq \epsilon_{\mathcal{Q},\mathcal{Z}} + \frac{1}{T}\sum_{t=0}^{T-1}\left\{\frac{56(T-t-1)^2 \log\{TN_{lj}\left(2\epsilon L_t, \mathcal{L}_{q_t(z'),z}, \|\cdot\|\right)/\delta\}}{3n}\right\} + o_p(n^{-1}).$$

where $L_t = KB_r(T-t-1)\Lambda$ for an absolute constant K.

3 Experiments

Please see the appendix for simulated experiments illustrating the benefits of orthogonal learning for estimation, and policy optimization. We apply our method to the MIMIC-III critical care database

for sepsis management, a task previously studied in the offline RL literature. Finally, in Figure 1a we summarize how the robust optimal actions change as the sensitivity parameter Λ is increased. We coarsen the 5×5 action space into four groups: no/less treatment (low action indices for both fluids/vasopressors), only IV fluid (high fluid action index), only vasopressors, and both fluid and vasopressors (high action indices for both fluids/vasopressors). At the far left, we have $\Lambda = 1$, which corresponds to the nominal policy, where there is an even mix of treatments. As Λ increases, the number of untreated or those receiving only fluid drops. Overall the robust policies move away from fluid-only actions towards vasopressors. This is in line with meta-analyses and studies in the clinical literature that suggest that conservative management (especially if concerned about mortality risk [5]) is aligned with preferring vasopressors to IV fluids[1, 4].

References

[1] P Marik and Rinaldo Bellomo. A rational approach to fluid therapy in sepsis. *BJA: British Journal of Anaesthesia*, 116(3):339–349, 2016.

[2] James M Robins, Andrea Rotnitzky, and Daniel O Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

[3] Paul R Rosenbaum. Design sensitivity in observational studies. *Biometrika*, 91(1):153–164, 2004.

[4] Matthew W Semler, David R Janz, Jonathan D Casey, Wesley H Self, and Todd W Rice. Conservative fluid management after sepsis resuscitation: a pilot randomized trial. *Journal of intensive care medicine*, 35(12):1374–1382, 2020.

[5] Jonathan A Silversides, Emmet Major, Andrew J Ferguson, Emma E Mann, Daniel F McAuley, John C Marshall, Bronagh Blackwood, and Eddy Fan. Conservative fluid management or deresuscitation for patients with sepsis or acute respiratory distress syndrome following the resuscitation phase of critical illness: a systematic review and meta-analysis. *Intensive care medicine*, 43(2):155–170, 2017.

[6] Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.