Theorem [theorem]Lemma [theorem]Proposition [theorem]Remark [theorem]Corollary [theorem]Definition

Robust Fitted-Q-Evaluation and Iteration under Sequentially Exogenous Unobserved Confounders

David Bruns-Smith BRUNS-SMITH@BERKELEY.EDU
Department of EECS
University of California
Berkeley, CA

Angela Zhou ZHOUA@USC.EDU
Department of Data Sciences and Operations
University of Southern California

Abstract

Offline reinforcement learning is important in domains such as medicine, economics, and e-commerce where online experimentation is costly, dangerous or unethical, and where the true model is unknown. We study robust policy evaluation and policy optimization in the presence of sequentially-exogenous unobserved confounders under a sensitivity model. We propose and analyze orthogonalized robust fitted-Q-iteration that uses closed-form solutions of the robust Bellman operator to derive a loss minimization problem for the robust Q function, and adds a bias-correction to quantile estimation. Our algorithm enjoys the computational ease of fitted-Q-iteration and statistical improvements (reduced dependence on quantile estimation error) from orthogonalization. We provide sample complexity bounds, insights, and show effectiveness both in simulations and on real-world longitudinal healthcare data of treating sepsis. In particular, our model of sequential unobserved confounders yields an online Markov decision process, rather than partially observed Markov decision process: we illustrate how this can enable warm-starting optimistic reinforcement learning algorithms with valid robust bounds from observational data.

We consider a finite-horizon Markov Decision Process on the full-information state space comprised of a tuple $\mathcal{M} = (\mathcal{S} \times \mathcal{U}, \mathcal{A}, R, P, \chi, T)$. (We consider the infinite horizon in the appendix). We let the state spaces $\mathcal{S}$,$\mathcal{U}$ be continuous, and to start assume the action space $\mathcal{A}$ is finite. The Markov decision process dynamics proceed from $t = 0, \ldots, T-1$ for a finite horizon of length T. (Although we focus on presenting the finite-horizon case, method and results extend readily to the discounted infinite-horizon case.) Let $\Delta(X)$ denote probability measures on a set X. The set of time t transition functions P is defined with elements $P_t : \mathcal{S} \times \mathcal{U} \times \mathcal{A} \to \Delta(\mathcal{S} \times \mathcal{U})$; R denotes the set of time t reward maps with $R_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$; the initial state distribution is $\chi \in \Delta(\mathcal{S} \times \mathcal{U})$. A policy, π, is a set of maps $\pi_t : \mathcal{S} \times \mathcal{U} \to \Delta(\mathcal{A})$, where $\pi_t(a \mid s, u)$ describes the probability of taking actions given states and unobserved confounders. Given the initial state distribution, the Markov Decision Process dynamics under policy π induce the random variables, for all t, $A_t \sim \pi_t(\cdot \mid S_t, U_t)$, $S_{t+1}, U_{t+1} \sim P_t(\cdot \mid S_t, U_t, A_t)$. When another type of norm is not indicated, we let $\|f\| := \mathbb{E}[f^2]^{1/2}$ indicate the 2-norm.

We consider a confounded offline setting: data is collected via an arbitrary behavior policy π^b that potentially depends on U_t, but in the resulting data set, the $\mathcal{U}$ part of the state space is unobserved.

As in standard offline RL, we study policy evaluation and optimization for target policies π^e using data collected under π^b. We will use P_π and $\mathbb{E}_\pi$ to denote the joint probabilities (and expectations thereof) of the random variables $S_t, U_t, A_t, \forall t$ in the underlying MDP running policy π. For the special case of the behavior policy π^b, we will write P_{obs}, $\mathbb{E}_{\text{obs}}$ to emphasize the distribution of variables in the observational dataset.

Our objects of interest will be the observed state Q function and value function for the target policy π^e:

$$\begin{aligned} Q_t^{\pi^e}(s,a) &:= \mathbb{E}_{\pi^e}[\textstyle\sum_{j=t}^{T-1} R(S_j, A_j, S_{j+1}) | S_t = s, A_t = a] \\ V_t^{\pi^e}(s) &:= \mathbb{E}_{\pi^e}[Q_t^{\pi^e}(S_t, A_t) | S_t = s]. \end{aligned} \tag{1}$$

We would like to find a policy π^e that is a function of the observed state alone, maximizing $V_t^{\pi^e}$. With unobserved confounders, we cannot directly evaluate the true expectations above due to biased estimation.

Assumption 1 (Memoryless unobserved confounders)**.** *The unobserved state U_{t+1} is independent of S_t, U_t, A_t.*

With memoryless unobserved confounders, observed-state policy evaluation and optimization in the full POMDP reduce to an MDP problem if only we knew the true marignal transitions. Define the marginal transition probabilities: $P_t(s_{t+1}|s_t, a_t) := \int_{\mathcal{U}} P_t(u_t|s_t) P_t(s_{t+1}|s_t, a_t, u_t) du_t$. Then we have the following proposition:

Proposition 1 (Marginal MDP)**.** *Given Assumption 1, for any policy π^e that is a function of S_t alone, the distribution of $S_t, A_t, \forall t$ in the full-information MDP running π^e is equivalent to the distribution of $S_t, A_t, \forall t$ in the marginal MDP, $(\mathcal{S}, \mathcal{A}, R, P, \chi, T)$. That is, $S_0 \sim \chi$, $A_t \sim \pi^e(\cdot \mid S_t)$, $S_{t+1} \sim P_t(\cdot|S_t, A_t)$.*

The key takeaway is that if we knew the true marginal transition probabilities, $P_t(S_{t+1}|S_t, A_t)$, then we could apply standard RL algorithms for evaluation or optimization. We have observed-state Q and value functions in the marginal MDP:

$$Q_t^{\pi^e}(s,a) = \mathbb{E}_{P_t}[R_t + Q_{t+1}^{\pi^e}(S_{t+1}, \pi_{t+1}^e) | S_t = s, A_t = a], \;\; V_t^{\pi^e}(s) = \mathbb{E}_{A \sim \pi_t^e(s)}[Q_t^{\pi^e}(s, A)]$$

Offline RL and Unobserved Confounding

Proposition 2 (Confounding for Regression)**.** *Let $f : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be any function. Given **??**, $\forall s, a$,*

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

This proposition shows that regression of f on states and actions using data collected according to π^b is a biased estimator.

Since the unobserved factor $\frac{\pi_t^b(A_t|S_t)}{\pi_t^b(A_t|S_t,U_t)}$ can be arbitrarily large without further assumptions, to make progress we follow the sensitivity analysis literature in causal inference.

Assumption 2 (Marginal Sensitivity Model)**.** *There exists* Λ *such that* $\forall t, s \in \mathcal{S}, u \in \mathcal{U}, a \in \mathcal{A}$*,*

$$\Lambda^{-1} \leq \left(\frac{\pi_t^b(a|s,u)}{1-\pi_t^b(a|s,u)}\right) / \left(\frac{\pi_t^b(a|s)}{1-\pi_t^b(a|s)}\right) \leq \Lambda. \tag{2}$$

The parameter Λ for this commonly-used sensitivity model in causal inference (Tan, 2012) has to be chosen with domain knowledge. Now consider any function $f : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. We can express the target expectation $\mathbb{E}_{P_t}[Y_t|S_t, A_t]$ as a weighted regression under the behavior policy with bounded weights. Define the random variable

$$W_t^{\pi^b} := \frac{\pi_t^b(A_t|S_t)}{\pi_t^b(A_t|S_t, U_t)}, \qquad \text{where } \mathbb{E}_{P_t}[Y_t|S_t, A_t] = \mathbb{E}_{\text{obs}}[W_t^{\pi^b} Y_t|S_t, A_t] \tag{3}$$

it satisfies the bounds

$$\alpha_t(S, A) \leq W_t^{\pi^b} \leq \beta_t(S, A), \forall s' \tag{4}$$
$$\alpha_t(S, A) := \pi_t^b(A_t|S_t) + \Lambda^{-1}(1 - \pi_t^b(A_t|S_t)), \quad \beta_t(S, A) := \pi_t^b(A_t|S_t) + \Lambda(1 - \pi_t^b(A_t|S_t)).$$

For each weight W_t that satisfies these constraints, there is a corresponding transition probability in the set:

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

Define the set $\mathcal{P}_t$ of transition probabilities for all s, a to be the product set over the $\mathcal{P}_t^{s,a}$. Then under Assumptions 1 and 2, the true marginal transition probabilities belong to $\mathcal{P}_t$. While point estimation is not possible, we can find the worst-case values of $Q_t^{\pi^e}$ and $V_t^{\pi^e}$ over transition probabilities in the uncertainty set, $\bar{P}_t \in \mathcal{P}_t$ — a Robust Markov Decision Process (RMDP) problem (Iyengar, 2005). Importantly, the set $\mathcal{P}_t$ is s, a-rectangular, and so we can use the results in Iyengar (2005) to define robust Bellman operators and a corresponding robust Bellman equation. Denote the robust Q and value functions $\bar{Q}_t^{\pi^e}$ and $\bar{V}_t^{\pi^e}$ and define the following operators:

Definition 1 (Robust Bellman Operators)**.** *For any function* $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$*,*

$$(\bar{\mathcal{T}}_t^{\pi^e} g)(s, a) := \inf_{\bar{P}_t \in \mathcal{P}_t} \mathbb{E}_{\bar{P}_t}[R_t + g(S_{t+1}, \pi_{t+1}^e)|S_t = s, A_t = a], \tag{5}$$
$$(\bar{\mathcal{T}}_t^{*} g)(s, a) := \inf_{\bar{P}_t \in \mathcal{P}_t} \mathbb{E}_{\bar{P}_t}[R_t + \max_{A'}\{g(S_{t+1}, A')\}|S_t = s, A_t = a]. \tag{6}$$

Proposition 3 (Robust Bellman Equation)**.** *Let* $|\mathcal{A}| = 2$ *and let Assumptions 1 and 2 hold. Then applying the results in Iyengar (2005), gives*

$$\bar{Q}_t^{\pi^e}(s, a) = \bar{\mathcal{T}}_t^{\pi^e} \bar{Q}_{t+1}^{\pi^e}(s, a), \quad \bar{V}_t^{\pi^e}(s) = \mathbb{E}_{A \sim \pi_t^e(s)}[\bar{Q}_t^{\pi^e}(s, A)],$$
$$\bar{Q}_t^{*}(s, a) = \bar{\mathcal{T}}_t^{*} \bar{Q}_{t+1}^{*}(s, a), \quad \bar{V}_t^{*}(s) = \mathbb{E}_{A \sim \bar{\pi}_t^*(s)}[\bar{Q}_t^{*}(s, A)],$$

where $\bar{Q}_t^*$ *and* $\bar{V}_t^*$ *are the optimal robust Q and value function achieved by the policy* $\bar{\pi}^*$*.*

Algorithm 1 Confounding-Robust Fitted-Q-Iteration

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 $\hat{\overline{Q}}_T = 0$.
for $t = T-1, \ldots, 1$ **do**
  Compute the nominal outcomes $\{Y_t^{(i)}(\hat{\overline{Q}}_{t+1})\}_{i=1}^n$ as in **??**.
  For $a \in \mathcal{A}$, fit $\hat{Z}_t^{1-\tau}$ the $(1-\tau)$th conditional quantile of the outcomes $Y_t^{(i)}$.
  Compute pseudooutcomes $\{\tilde{Y}_t^{(i)}(\hat{Z}_t^{1-\tau}, \hat{\overline{Q}}_{t+1})\}_{i=1}^n$ as in **??**.
  For $a \in \mathcal{A}$, fit $\hat{\overline{Q}}_t$ via least-squares regression of $\tilde{Y}_t^{(i)}$ against $(S_t^{(i)}, A_t^{(i)})$.
  Compute $\pi_t^*(s) \in \arg\max_a \hat{\overline{Q}}_t(s,a)$.
end for

Method Nominal (non-robust) FQI (Ernst et al., 2006; Le et al., 2019; Duan et al., 2021) successively forms approximations $\hat{Q}_t$ at each time step by minimizing the Bellman error. In our robust version of FQI, we instead approximate the robust Bellman operator with function approximation.

Proposition 4. *Let Q be a real-valued function over states and actions, and define $Y_t(Q)$ the Bellman target. The robust $Q(s,a)$ function solves the following optimization problem:*

$$(\bar{\mathcal{T}}_t^* Q)(s,a) = \min_{W_t} \big\{ \mathbb{E}_{obs}\left[W_t Y_t(Q) | S_t = s, A_t = a\right] : \\ \mathbb{E}_{obs}\left[W_t | S_t = s, A_t = a\right] = 1, \;\; \alpha_t(S,A) \leq W_t \leq \beta_t(S,A), a.e. \big\}.$$

The closed-form state-action conditional solution to **??** is written in terms of a superquantile (also called conditional expected shortfall, or covariate-conditional CVaR). The conditional expected shortfall is the conditional expectation of exceedances of a random variable beyond its conditional quantile. Define $\tau \coloneqq \Lambda/(1+\Lambda)$. For any function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define the observational $(1-\tau)$-level conditional quantile of the Bellman target:

$$Z_t^{1-\tau}(Y_t(Q) \mid s,a) \coloneqq \inf_z \{z \colon P_{\text{obs}}(Y_t(Q) \geq z \mid S_t = s, A_t = a) \leq 1-\tau\}.$$

We use the following shorthands when clear from context: $Z_{t,a}^{1-\tau} \coloneqq Z_t^{1-\tau}(Y_t(Q) \mid s,a), \alpha_t \coloneqq \alpha_t(S,A), \beta_t \coloneqq \beta_t(S,A)$.

Proposition 5. *The solution to the robust Bellman operator is:*

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

To avoid transferring biased first-stage estimation error of $Z_t^{1-\tau}$ to the Q-function, we apply an orthogonalization of Olma (2021) to obtain our regression target for robust FQE:

$$\tilde{Y}_t(Z,Q) \coloneqq \alpha_t Y_t(Q) + \tfrac{1-\alpha_t}{1-\tau}\Big(Y_t(Q)\mathbb{I}\left[Y_t(Q) \leq Z_t^{1-\tau}\right] - Z \cdot \{\mathbb{I}\left[Y_t(Q) \leq Z\right] - (1-\tau)\}\Big) \tag{8}$$

When the quantile functions are consistent, the orthogonalized pseudo-outcome enjoys quadratic, not linear on the first-stage estimation error in the quantile functions. We describe in more detail in the next section on guarantees. The orthogonalized time-t target of estimation is:

$$\hat{\overline{Q}}_t \in \arg\min_{q_t} \mathbb{E}_{n,t}[(\tilde{Y}_t(\hat{Z}_t^{1-\tau}, \hat{\overline{Q}}_{t+1}) - q_t(S_t, A_t))^2]. \tag{9}$$

Guarantees

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

$$\|\hat{\overline{Q}}_t(S,a) - \overline{Q}_t(S,a)\| \leq \|\tilde{\overline{Q}}_t(S,a) - \overline{Q}_t(S,a)\| + o_p(n^{-\frac{1}{2}}).$$

Theorem 1 (Fitted Q Iteration guarantee)**.** *Suppose C-concentratability and ϵapproximate Bellman completeness and let B_R be the bound on rewards. Recall that $\mathcal{E}(\hat{Q}) = \frac{1}{T}\sum_{t=0}^{T-1} \left\| \hat{Q}_t - \overline{\mathcal{T}}_t^\star \hat{Q}_{t+1} \right\|_{\mu_t}^2$. Then, with probability $> 1-\delta$, under assumption of a finite function class, we have that*

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

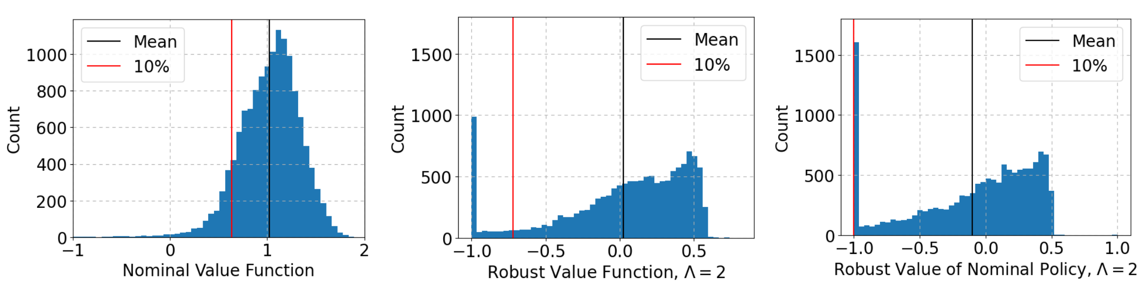
while under an infinite function class with bracketing numbers, choosing the covering number approximation error $\epsilon = O(n^{-1})$ such that $\epsilon_{\mathcal{Q},\mathcal{Z}} = O(n^{-1})$, we have that

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

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

1. Experiments

See the appendix/full paper for details.



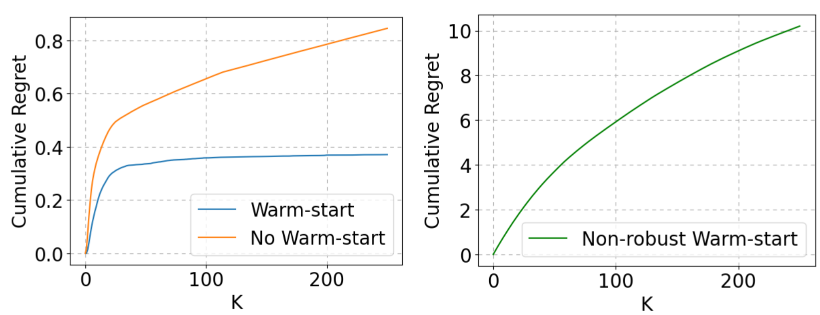
(a) (b) (c)

Figure 1: Histograms of initial state value functions over the observed initial states in the MIMIC-III dataset. From left to right, the nominal value; the robust value for $\Lambda = 2$; and the robust value of the nominal optimal policy for $\Lambda = 2$.

Extension: Warmstarting We can use our robust valid bounds to warm-start online algorithms via valid robust bounds from observational data. See the appendix for details.

Λ	Algorithm	MSE($\bar{V}_0^*$)	ℓ_2 Parameter Error	% wrong action
1	FQI	0.2300	3.399	28%
2	Non-Orthogonal	0.5496	4.057	31%
	Orthogonal	0.5271	3.522	28%
5.25	Non-Orthogonal	3.160	11.51	43%
	Orthogonal	1.739	3.949	31%
8.5	Non-Orthogonal	7.683	24.04	45%
	Orthogonal	2.723	3.921	31%
11.75	Non-Orthogonal	15.22	48.89	47%
	Orthogonal	3.397	3.725	31%
15	Non-Orthogonal	30.21	88.02	48%
	Orthogonal	3.848	3.462	30%

Table 1: Simulation results with $d = 100$ and $n = 600$, reporting the value function MSE, Q function parameter error, and the portion of the time a sub-optimal action is taken. The results compare non-orthogonal and orthogonal confounding robust FQI over five values of Λ.



(a) (b)

Figure 2: Simulation results for online LSVI-UCB. Panel (a) plots the cumulative regret of LSVI-UCB without warm-starting, and with robust warm-starting. Panel (b) plots the cum. regret of LSVI-UCB where the offline data is naively treated as if had been collected online.

References

Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. In International Conference on Machine Learning, pages 2892–2902. PMLR, 2021.

Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In Proceedings of the 45th IEEE Conference on Decision and Control, pages 667–672. IEEE, 2006.

Garud N Iyengar. Robust dynamic programming. Mathematics of Operations Research, 30 (2):257–280, 2005.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In International Conference on Machine Learning, pages 3703–3712. PMLR, 2019.

Tomasz Olma. Nonparametric estimation of truncated conditional expectation functions. arXiv preprint arXiv:2109.06150, 2021.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. Journal of the American Statistical Associatioon, 2012.