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

David Bruns-Smith
University of California Berkeley
bruns-smith@berkeley.edu

Angela Zhou
Data Sciences and Operations
University of Southern California
zhoua@usc.edu

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. However, most methods assume all covariates used in the behavior policy's action decisions are observed. This untestable assumption may be incorrect. 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. 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 in simulations.

1 Introduction

Sequential decision-making problems in medicine, economics, and e-commerce require the use of historical observational data when online experimentation is costly, dangerous or unethical. Given the rise of big data, there is great potential to improve decisions based on personalizing treatments to those who most benefit. However, it is also more difficult to ex-ante specify the underlying dynamics when personalizing sequential decision-making from rich data, which precludes performance evaluation via traditional methods based on stochastic simulation. The recent literature on offline reinforcement learning addresses these challenges of evaluating sequential decision rules, given only a historical dataset of observed trajectories, for example methods that target estimation of the Q function leveraging black-box regression.

However, these methods almost unilaterally all assume full observability of all the covariate information that informed historical treatment decisions. Unfortunately, historical decision-making policies typically made decisions based on additional unobserved variables. Such data was usually collected for convenience from "business as usual", i.e. neither a randomized controlled trial nor a carefully designed observational cohort study, typically arises from a system that was optimizing for outcomes, or other complex human decisions. This introduces *unobserved confounders*, variables that impact both treatment assignment and outcomes. In the presence of unmeasured confounders, the typical approach of estimating transition probabilities and solving standard Markov Decision Processes is biased due to incomplete adjustment for confounding.

The default realistic case for observational data is that there were some unobserved confounders; but their influence can be limited as data becomes richer in the era of big data. For example, if working with a database of electronic health records, it corresponds to assuming that most of medical decision-making can be explained by information such as recorded patient vitals, while unobserved confounders such as patient affect were less important in medical-decision making. *Sensitivity analysis* techniques in the causal inference literature assess the impact of potential unobserved confounding. Instead of reporting incorrect point estimates, they report the range of estimates consistent with some potential amount of unobserved confounding, via how it affects the probability of selection into treatment [Robins et al., 2000, Rosenbaum, 2004, VanderWeele and Ding, 2017]. These estimates can be framed as optimization problems over ambiguity sets, which can be sized by domain expertise, for example by comparing to the informativity of observed covariates. Importantly, such restrictions on the unobserved confounding are untestable from observational data, and ambiguity sets on unobserved confounding differ from uncertainty sets motivated on probabilistic grounds alone, i.e. robustness to finite-sample deviations.

We study robust sequential personalized policy learning under an ambiguity set of the unknown probability of taking actions given *both* observed and unobserved confounders, the *propensity score*. We seek not only robust bounds on value, but also robust decisions. Our algorithm links sensitivity analysis under unobserved confounders to the framework of robust Markov decision processes, and uses statistical function approximation to estimate bounds on the worst-case conditional bias of the Q function. More specifically, we use the "marginal sensitivity model" (MSM) of Tan [2012], a variant of Rosenbaum's sensitivity model [Rosenbaum, 2004], which has been widely used for offline single-timestep policy optimization [Aronow and Lee, 2013, Miratrix et al., 2018, Zhao et al., 2019, Yadlowsky et al., 2018, Kallus et al., 2018, Kallus and Zhou, 2020b]. Contrary to typical uses of the MSM probing importance sampling-based estimators, we partially identify the Bellman equation for the state-action value function using an MSM with state-conditional restrictions. We develop the first principled and practical methodology for robust sequential policy learning under memoryless unobserved confounders. Recent work has only solved robust policy evaluation (not learning) under the sequential MSM under restrictions such as one-stage unobserved confounders [Namkoong et al., 2020], or small, discrete state spaces under additional assumptions [Kallus and Zhou, 2020a, Bruns-Smith, 2021]. Partially identifying the Bellman equation provides a direct connection to practical policy optimization algorithms such as the fitted-Q-iteration we extend.

Learning from observational data is crucial to make progress on data-driven decision-making in consequential domains where online reinforcement learning is infeasible or costly. For example, the release of electronic health records such as the MIMIC-III critical care database enabled rich data-driven research on medical decision-making: researchers developed an illustrative task for offline reinforcement learning based on managing sepsis via administration of vasopressors and fluids, a complex dynamic task without clinical consensus. This is an important problem: sepsis is one of the foremost drivers of both mortality and hospital costs. But in the causal and reinforcement learning setting, typical performance measures in machine learning such as cross-validation, or simulation of sequential policies using a known generative model are *not valid*. Instead, the performance evaluation of learned sequential decision policies via off-policy evaluation from offline reinforcement learning implicitly requires the untrue assumption of unconfoundedness. Gottesman et al. [2019] thoroughly articulates these challenges of offline evaluation, including the likely presence of no unobserved confounders in this dataset. Importantly, such real-world data is complex, motivating scalable approaches based on statistical learning for generalization to unseen states. Robust evaluation can enable learning from observational data.

In this paper, we develop methodology for robust bounds and decision rules that can inform managerial decisions in a number of ways. Later on, we revisit sepsis data from MIMIC-III: since our method allows direct comparison to typical fitted-Q-evaluation/iteration methods used in the literature, we show how comparing robust vs. nominal value functions can provide insight or inform future investigation. More broadly, the FDA has recognized a growing need for methods that assess the "robustness and resilience of these [clinical decision support] algorithms to withstand changing clinical inputs and conditions" [FDA, 2021]. A recent working group argues that sensitivity analysis can support product development from real-world evidence and points out the need for comparable methodology for the sequential policy learning setting [Ding et al., 2023]. Finally, even if robust policies are not deployed directly, robust bounds can be used as prior knowledge to improve the data-efficiency of online experimentation, if it becomes available. We introduce an extension of our methods for warm-starting online reinforcement learning, which also highlights key differences of our structural assumptions from other models for Markov Decision Processes with unobserved confounders: the online counterpart under our structural assumption of no memoryless unobserved confounders is a tractable MDP instead of an intractable partially observable Markov decision process (POMDP).

Contributions: we develop an algorithm for *efficiently computing* MSM bounds with multi-step confounding, high-dimensional continuous state spaces and function approximation. Our approach leverages the recent characterization of sensitivity in single-step settings as a conditional expected shortfall (also called conditional CVaR or superquantile) [Shapiro et al., 2021]. Our algorithm is a simple extension of fitted-Q evaluation/iteration [Fu et al., 2021, Le et al., 2019] that can be implemented with off-the-shelf supervised learning algorithms, making it easily accessible to practitioners. We solve a key *statistical* challenge by incorporating orthogonalized estimation of the robust Bellman operator, and derive a corresponding *theoretical* analysis, giving sample complexity guarantees for orthogonalized robust FQI based on the richness of the approximating function classes. This reduces the dependence of statistical error in estimating the conditional expected shortfall on estimation of the conditional quantile function. Finally, we show how our model enables warm-starting standard online optimistic reinforcement learning from valid robust bounds for safe data-efficiency. Our algorithm enables researchers in the managerial, clinical, and social sciences to assess and report sensitivity to unobserved confounding for dynamic policies learned from observational data, and to learn new policies that are more robust when assumptions on confounders fail.

2 Related Work

We first discuss offline reinforcement learning in general, and other approaches for unobserved confounders besides ours based on robustness. Then we discuss other topics such as orthogonalized estimation, robust Markov decision processes, and robust offline reinforcement learning; before summarizing how our work is at the intersection of and relates to these areas.

Policy learning with unobserved confounders in single-timestep and sequential settings. The rapidly growing literature on offline reinforcement learning with unobserved confounders can broadly be divided into three categories. We briefly discuss central differences from our approach to these three broad groups and include an expanded discussion in the appendix. First, some work assumes point identification is available via instrumental variables [Wang et al., 2021]/latent variable models [Bennett and Kallus, 2019]/front-door identification [Shi et al., 2022b]. Although point identification is nice if available, sensitivity analysis can be used when assumptions of point identification (instrumental-variables, front-door adjustment) *are not true*, as may be the case in practice. Second, a growing literature considers proximal causal inference in POMDPs from temporal structure [Tennenholtz et al., 2019, Bennett et al., 2021, Uehara et al., 2022, Shi et al., 2022a] or additional proxies [Miao et al., 2022]. Proximal causal inference imposes additional (unverifiable) completeness assumptions on the latent variable structure and is a statistically challenging ill-posed inverse problem. Furthermore, we study a more restricted model of memoryless unobserved confounders that precisely delineates unobserved confounding from general POMDP concerns. As a result, we have an online counterpart that is a marginal MDP, justifying warm-starting approaches. Third, a few approaches compute no-information partial identification (PI) bounds based only on the structure of probability distributions and no more. Han [2022] obtains a partial order on decision rules with only the law of total probability. [Chen and Zhang, 2021] derives PI bounds with time-varying instrumental variables, based on Manski-Pepper bounds. These can generally be much more conservative than sensitivity analysis, which relaxes strong assumptions.

Overall, developing a *variety* of identification approaches further is crucial both for analysts to use appropriate estimators/bounds, and methodologically to support falsifiability analyses. Other works include [Fu et al., 2022, Liao et al., 2021, Saghafian, 2021]. In our work, we consider the marginal sensitivity model. Extending to other sensitivity analysis models may also be of interest [Robins et al., 2000, Scharfstein et al., 2018, Yang and Lok, 2018, Bonvini and Kennedy, 2021, Bonvini et al., 2022, Scharfstein et al., 2021, Chernozhukov et al., 2022]. Both the state-action conditional uncertainty sets and the assumption of memoryless unobserved confounders are particularly crucial in granting state-action rectangularity (for binary treatments), and avoiding decision-theoretic issues with time-inconsistent preferences in multi-stage robust optimization [Delage and Iancu, 2015]. On the other hand, the exact functional form (subject to these structural assumptions) could readily be modified.

Recent work of Panaganti et al. [2022] also proposes a robust fitted-Q-iteration algorithm for RMDPs. Although the broad algorithmic design is similar, we consider a different uncertainty set from their ℓ_1 set, and further introduce orthogonalization. In the single-timestep setting, further improvements are possible when targeting a simpler scalar mean, such as in Dorn and Guo [2022], Dorn et al. [2021]. By constrast, we need to estimate the *entire robust Q-function.*

Off-policy evaluation in offline reinforcement learning An extensive line of work on off-policy evaluation [Jiang and Li, 2016, Thomas et al., 2015, Liu et al., 2018, Tang et al., 2019] in offline reinforcement learning studies estimating the policy value of a posited evaluation policy when only data from the behavior policy is available. Most of this literature, implicitly or explicitly, assumes sequential ignorability/sequential unconfoundedness. Methods for policy optimization are also different in the offline setting than in the online setting. Options include direct policy search (which is quite sensitive to functional specification of the optimal policy) [Zhao et al., 2015], off-policy policy gradients which are either statistically noisy [Imani et al., 2018] or statistically debiased but computationally inefficient [Kallus and Uehara, 2020b], or fitted-Q-iteration [Le et al., 2019, Ernst et al., 2006]. Of these, fitted-Q-iteration's ease of use and scalability make it a popular choice in practice. It is also theoretically well-studied [Duan et al., 2021]. A marginal MDP also appears in Kallus and Zhou [2022] but in a different context, without unobserved confounders.

Orthogonalized estimation. Double/debiased machine learning seeks so-called Neyman-orthogonalized estimators of statistical functionals so that the Gateaux derivative of the statistical functional with respect to nuisance estimators is 0 [Newey, 1994, Chernozhukov et al., 2018, Foster and Syrgkanis, 2019]. Nuisance estimators are intermediate regression steps (i.e. the conditional quantile) that are not the actual target function of interest (i.e. the robust Q function). Orthogonalized estimation reduces the dependence of the statistical estimator on the estimation rate of the nuisance estimator. See Kennedy [2022] for tutorial discussion and Jordan et al. [2022] for a computationally-minded tutorial. There is extensive literature on double robustness/semiparametric estimation in the longitudinal setting, often from biostatistics and statistics [Laan and Robins, 2003, Robins et al., 2000, Orellana et al., 2010]. Many recent works have studied double/debiased machine learning in the sequential and off-policy setting [Bibaut et al., 2019, Kallus and Uehara, 2020a, Singh and Syrgkanis, 2022, Lewis and Syrgkanis, 2020].

Recent work studies orthogonality/efficiency for partial identification and in other sensitivity models than the one here [Bonvini and Kennedy, 2021, Bonvini et al., 2022, Scharfstein et al., 2021, Chernozhukov et al., 2022]. Semenova [2017], Olma [2021] study orthogonalization of partial identification or conditional expected shortfall, and we build on some of their analysis in this paper. In particular, we directly apply the orthogonalization given in Olma [2021]. Yadlowsky et al. [2018] study orthogonality under the closely related Rosenbaum model and provide very nice theoretical results. They obtain their orthogonalization via a variational characterization of expectiles. Though Namkoong et al. [2020] consider a restricted model of the *worst* single-timestep confounding, out of all timesteps, it seems likely that sequential orthogonalization under the sequential exogenous confounders assumption is also possible. The single-timestep work of Jeong and Namkoong [2020] orthogonalizes a marginal CVaR, but they assume the quantile function is known. Dorn and Guo [2022] provide very nice and strong theoretical guarantees and surface additional properties of double validity. In Appendix C.3 we briefly highlight differences in estimating the marginal policy value rather than recovering the entire state-action conditional Q-function, as we require for policy optimization in this work. (Recovering the entire robust Q function is important for rectangularity). Hence estimation in this setting is qualitatively different from estimating the policy value.

Robust Markov Decision Processes and offline reinforcement learning. Elsewhere, in the robust Markov-decision process framework [Nilim and El Ghaoui, 2005], the challenge of *rectangularity* has been classically recognized as an obstacle to efficient algorithms although special models may admit non-rectangularity and computational tractability [Goyal and Grand-Clement, 2022]. The computational challenges of robust MDPs have been widely recognized due to requiring the solution of a robust optimization problem for evaluation; recent algorithmic improvements are typically tailored for special structure of ambiguity sets. On the other hand, work in robust Markov decision processes has prominently featured the role of uncertainty sets and coherent risk measures, for example in distributionally robust Markov Decision Processes [Zhou et al., 2021]. Our work relates sensitivity analysis in sequential causal inference to this line of literature and focuses on algorithms for policy evaluation based on a robust fitted-Q-iteration. Other relevant works include Lobo et al. [2020], which considers a "soft-robust" criterion that averages the nominal expectation and the robust expectation; however, they study *marginal* CVaR while our later discussion of CVaR is conditional. Studying the conditional expected shortfall (equivalently, CVaR) uncertainty set is a crucial difference from previous work on risk-sensitive MDPs [Chow et al., 2015]. More generally, the so-called "pessimism" principle in offline reinforcement learning is well-studied as a tool to relax strong concentratability assumptions [Jin et al., 2021]. Regarding distributionally robust offline reinforcement learning specifically, Ma et al. [2022] studies linear function approximation. Yang et al. [2022] studies the sample complexity of tabular robust MDPs under a generative model. The focus of our work is on unobserved confounders, although we reformulate the ambiguity set as a distributionally robust optimization problem. Other, less related, works study distributionally robust online learning [Wang et al., 2023].

Summary of differences of our work. We connect robustness for causal inference under unobserved confounders to distributionally robust MDPs and orthogonalized estimation, to obtain scalable methods with provable guarantees. In contrast to the line of work developing specialized (first-order) algorithms for (robust) Markov decision-processes, we consider approximate (robust) Bellman operator evaluations in the fitted-Q-evaluation/iteration paradigm. We use the closed-form characterization of the state-conditional solution to derive the infinite-data solution and approximate the estimation of the resulting function from data. Also, methodologically, we leverage orthogonalized estimation, which does not appear in previously mentioned works on distributionally robust offline reinforcement learning and can be of interest beyond our setting of unobserved confounders.

3 Problem Setup and Characterization

3.1 Problem Setup with Unobserved State

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 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*. That is, although the underlying dynamics follow a standard Markov decision process generating the history $\{(S_t^{(i)}, U_t^{(i)}, A_t^{(i)}, S_{t+1}^{(i)})_{t=0}^{T-1}\}_{i=1}^n$, the observational dataset omits the unobserved confounder. The observational dataset comprises of N trajectories including observed confounders only, $\mathcal{D}_{obs} := \{(S_t^{(i)}, A_t^{(i)}, S_{t+1}^{(i)})_{t=0}^{T-1}\}_{i=1}^n$ For example, we might have a data set of electronic medical records and treatment decisions made by doctors; the electronic medical records include an observed set of patient measurements S_t, but the doctors may have made their treatment decisions using additional unrecorded information U_t.

As in standard offline RL, we study policy evaluation and optimization for target policies π^e using data collected under π^b. In our confounded setting, we consider π^e that are a function of the observed state S_t alone.[1] 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:

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

We would like to find a policy π^e that is a function of the observed state alone, maximizing $V_t^{\pi^e}$. Throughout, we work primarily in the offline reinforcement learning setting where we do not have access to online exploration due to cost or safety concerns. With unobserved confounders, we cannot directly evaluate the true expectations above due to biased estimation. Therefore in the remainder of Section 3, we introduce confounding-*robust* Q and value functions, which we can estimate from the observational data.

3.2 Defining an MDP on Observables

We next articulate the challenges of our setting more specifically and introduce our main structural assumption of memoryless unobserved confounders. For offline policy evaluation/optimization with unobserved confounding, there are two separate concerns: biased estimation from confounded observational data, and partial observability in the presence of unobserved confounders. First, the dependence of π^b on U_t introduces unobserved confounding, sothe distribution of the observed data is biased for estimating the true underlying transition probabilities. Without further assumptions, the observational distribution alone cannot completely adjust for the spurious correlation induced by the behavior policy. Second, even if we knew the true underlying transition probabilities, the existence of the unobserved state would change the policy optimization problem from a tractable MDP to an intractable Partially-Observed Markov Decision Process (POMDP). Standard RL algorithms like Bellman iteration for MDPs would no longer yield an optimal policy — because, for example, the observed next state S_{t+1} need not be Markovian conditional on only S_t and A_t.

In this section, we *isolate* the confounding concern from the POMDP concern by introducing a "memoryless confounding" assumption. Under this assumption, we will show that policy evaluation over π^e in the underlying MDP is equivalent to policy evaluation in a marginal MDP over the observed state alone. Therefore, the underlying difficulty of decision-making under memoryless unobserved confounders is intermediate between the unconfounded and generic POMDP setting.

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

[1] Note that in our setup, the practitioner specifies the reward as a function of only the observed state $\mathcal{S}$. This is essentially without loss of generality since the reward depends on S_{t+1} and S_{t+1} depends on U_t.

Under this assumption, the full-information transition probabilities factorize as:

$$\begin{aligned} P_t(s_{t+1}, u_{t+1}|s_t, a_t, u_t) &= P_t(s_{t+1}|s_t, a_t, u_t)P_t(u_{t+1}|s_{t+1}, s_t, a_t, u_t) \\ &= \underbrace{P_t(s_{t+1}|s_t, a_t, u_t)}_{\text{new observed state}} \underbrace{P_t(u_{t+1}|s_{t+1})}_{\text{new unobserved state}} . \end{aligned}$$

In a slight abuse of notation, we will change the subscript on the unobserved state distribution to read $P_{t+1}(u_{t+1}|s_{t+1})$ so that the time subscripts are consistent. Note that under Assumption 1, $P_{\text{obs}}(U_t|S_t)$ is always the same regardless of what policy produced the historical data. Without Assumption 1, $P_{\text{obs}}(U_t|S_t)$ would generally vary with the behavior policy π^b because U_t could depend on S_{t-1}, A_{t-1}, and U_{t-1}.

With memoryless unobserved confounders, observed-state policy evaluation and optimization in the full POMDP reduce to an MDP problem. Define the marginal transition probabilities:

$$P_t(s_{t+1}|s_t, a_t) \coloneqq \int_{\mathcal{U}} P_t(u_t|s_t)P_t(s_{t+1}|s_t, a_t, u_t)du_t \tag{2}$$

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, $S_0 \sim \chi$, $A_t \sim \pi^e(\cdot \mid S_t)$, $S_{t+1} \sim P_t(\cdot|S_t, A_t)$.*

See the Appendix for a formal derivation. The key takeaway from Proposition 1 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, that satisfy the Bellman evaluation equations,

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

where we use the short-hands $R_t \coloneqq R_t(S_t, A_t, S_{t+1})$ and $g(S', \pi) \coloneqq \mathbb{E}_{A'\sim\pi(S')}[g(S', A')]$ for any $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Furthermore, by classical results [Puterman, 2014], an optimal policy exists among policies defined on the observed state alone, yielding the optimal Q function, $Q_t^*(s, a)$, and value function, $V_t^*(s)$, with corresponding Bellman optimality equations.

Before continuing, we want to emphasize that while Assumption 1 is strong, it has testable implications. In particular, under Assumption 1 the observed-state transition probabilities will be *Markovian*, which can be tested from observed states and actions alone.[2]

3.3 Offline RL and Unobserved Confounding

Proposition 1 establishes that the oracle decision problem, given knowledge of the true marginal transition probabilities, remains a Markov Decision Process under memoryless confounding. However, while Assumption 1 rules out POMDP concerns, it does not rule out bias from unobserved confounding. In general, it is *not* possible to get unbiased estimates of the true marginal

[2]It is possible to use observed-state Markovian transitions as the core assumption at the cost of substantially more complexity. See the Appendix for discussion.

observed-state transitions given data collected under π^b when U_t is unobserved. In particular, $P_{\text{obs}}(S_{t+1}|S_t, A_t) \neq P_t(S_{t+1}|S_t, A_t)$. To see this, first define the marginal behavior policy,

$$\pi_t^b(a_t|s_t) := \int_{\mathcal{U}} \pi_t^b(a_t|s_t, u_t) P_t(u_t|s_t) du_t = P_{\text{obs}}(a_t|s_t).$$

Then,

$$\begin{aligned} P_{\text{obs}}(s_{t+1}|s_t, a_t) &= \int_{\mathcal{U}} P_t(s_{t+1}|s_t, a_t, u_t) P_{\text{obs}}(u_t|s_t, a_t) du_t \\ &= \int_{\mathcal{U}} P_t(s_{t+1}|s_t, a_t, u_t) \frac{\pi_t^b(a_t|s_t, u_t)}{\pi_t^b(a_t|s_t)} P_t(u_t|s_t) du_t, \end{aligned} \tag{3}$$

where the second equality follows by Bayes rule. The final expression for $P_{\text{obs}}(s_{t+1}|s_t, a_t)$ differs from $P_t(s_{t+1}|s_t, a_t)$ in eq. (2) by the unobserved factor $\frac{\pi_t^b(a_t|s_t,u_t)}{\pi_t^b(a_t|s_t)}$. Note that the term $P_{\text{obs}}(u_t|s_t, a_t)$ is the bias from confounding: in the observational distribution conditioning on a_t changes the distribution of the unobserved u_t relative to $P_t(u_t|s_t)$ because a_t is drawn according to $\pi^b(a_t|s_t, u_t)$.

If π^b is independent of u_t, the ratio $\frac{\pi_t^b(a_t|s_t,u_t)}{\pi_t^b(a_t|s_t)}$ will be uniformly 1 and we recover $P_t(s_{t+1}|s_t, a_t)$. However, if $\pi_t^b(a_t|s_t, u_t)$ can be arbitrary, then an estimate of $P_t(s_{t+1}|s_t, a_t)$ using $P_{\text{obs}}(s_{t+1}|s_t, a_t)$ can be arbitrarily biased. This result immediately implies that any regression using P_{obs} will be biased for the corresponding estimand in the marginal MDP.

Proposition 2 (Confounding for Regression)**.** *Let $f : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be any function. Given Assumption 1, $\forall s, a$,*

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

where the first equality follows from Proposition 1 and the second equality follows from Equation (3). This proposition shows that regression of f on states and actions using data collected according to π^b is a biased estimator for the corresponding conditional expectation under the true marginal transition probabilities $P_t(s'|s, a)$ where the exact bias is:

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

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 $\forall t, s \in \mathcal{S}, u \in \mathcal{U}, a \in \mathcal{A}$,*

$$\Lambda^{-1} \leq \left(\frac{\pi_t^b(a \mid s, u)}{1 - \pi_t^b(a \mid s, u)}\right) / \left(\frac{\pi_t^b(a \mid s)}{1 - \pi_t^b(a \mid s)}\right) \leq \Lambda. \tag{4}$$

The parameter Λ for this commonly-used sensitivity model in causal inference [Tan, 2012] has to be chosen with domain knowledge. A common approach is to compare Λ to corresponding values for observed variables, e.g. in a clinical setting, if smoking has an effective $\Lambda = 1.5$, a practitioner might say "I do not believe there exists an unobserved variable with twice the explanatory power of smoking" to justify a choice of $\Lambda = 3$ [Hsu and Small, 2013].

Now consider any function $f : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ as in Proposition 2. For shorthand, we will write $Y_t := f(S_t, A_t, S_{t+1})$. We use a generic f here to emphasize that this argument would apply to any model-based or model-free RL algorithms using regression, but later when we introduce our fitted-Q iteration algorithm, we will specialize Y_t to get an empirical estimate of the Bellman operator. Combining Assumption 2 and Proposition 2, 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

$$W_t^{\pi^b} := \frac{\pi_t^b(A_t|S_t)}{\pi_t^b(A_t|S_t, U_t)}, \qquad \text{where } \mathbb{E}_{P_t}[Y_t|S_t, A_t] = \mathbb{E}_{\text{obs}}[W_t^{\pi^b} Y_t|S_t, A_t] \qquad \text{(proposition 2).} \tag{5}$$

While we cannot estimate $W_t^{\pi^b}$, we can bound it. The weights must satisfy the density constraint

$$\mathbb{E}_{\text{obs}}[W_t^{\pi^b}|S_t, A_t] = 1, \tag{6}$$

and Assumption 2 implies the following bounds almost everywhere:

$$\alpha_t(S, A) \leq W_t^{\pi^b} \leq \beta_t(S, A), \forall s' \tag{7}$$
$$\alpha_t(S, A) := \pi_t^b(A_t|S_t) + \Lambda^{-1}(1 - \pi_t^b(A_t|S_t)), \quad \beta_t(S, A) := \pi_t^b(A_t|S_t) + \Lambda(1 - \pi_t^b(A_t|S_t)).$$

So while Proposition 2 demonstrates that we cannot unbiasedly estimate the value function in the confounded setting, we can instead compute worst-case bounds on the conditional bias subject to the constraints in eqs. (6) and (7). Next, we will make this precise by showing that Assumption 2 defines a Robust Markov Decision Process.

3.4 Robust Estimands and Bellman Operators

In this section, we introduce our key estimands – the robust Q and value functions. Assumption 2 implies the constraints in eqs. (6) and (7), which define an uncertainty set for the true observed-state transition probabilities $P_t(s'|s, a)$. Kallus and Zhou [2020a] and Bruns-Smith [2021] uses a reparameterization to show that for each weight W_t that satisfies these constraints, there is a corresponding transition probability in the set:

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

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 $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,*

$$(\bar{\mathcal{T}}_t^{\pi^e} g)(s,a) \coloneqq \inf_{\bar{P}_t \in \mathcal{P}_t} \mathbb{E}_{\bar{P}_t}[R_t + g(S_{t+1}, \pi_{t+1}^e) | S_t = s, A_t = a], \tag{8}$$

$$(\bar{\mathcal{T}}_t^{*} g)(s,a) \coloneqq \inf_{\bar{P}_t \in \mathcal{P}_t} \mathbb{E}_{\bar{P}_t}[R_t + \max_{A'}\{g(S_{t+1}, A')\} | S_t = s, A_t = a]. \tag{9}$$

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 $\bar{\pi}^*$.

Finally, we comment on the tightness of the robust operator. For a fixed s and a, the $\mathcal{P}_t^{s,a}$ is exactly the set of transition probabilities consistent with Assumption 2 and the observational data distribution. However the s, a-rectangular product set $\mathcal{P}_t$ does not explicitly enforce the density constraint on π_t^b *across actions*, and is therefore potentially loose. In the special case where there are only two actions, Dorn et al. [2021] show that the different minima over $\mathcal{P}_t^{s,a}$ across actions are *simultaneously achievable*, and thus the robust bounds are tight and we get equalities in Proposition 3. For $|\mathcal{A}| > 2$, the infimum in eq. (8) is *not* generally simultaneously realizable (see Appendix A.1 for a counter-example). Nonetheless, the robust Bellman operator corresponds to an s, a-rectangular relaxation of the RMDP, Proposition 3 will hold with lower bounds instead of equalities, and our results are still guaranteed to be robust.

4 Method

In the previous section, we defined our estimands of interest — the robust Q and value functions under the marginal sensitivity model. In this section, we introduce robust policy optimization via function approximation. Our estimation strategy is a robust analog of Fitted-Q Iteration (FQI).

Assume that we observe n trajectories of length T, where the observational dataset $\mathcal{D}_{obs} \coloneqq \{(S_t^{(i)}, A_t^{(i)}, S_{t+1}^{(i)})_{t=0}^{T-1}\}_{i=1}^n$ was collected from the underlying MDP under an unknown behavior policy π^b that depends on the unobserved state. We will write $\mathbb{E}_{n,t}$ to denote a sample average of the n data points collected at time t, e.g. $\mathbb{E}_{n,t}[f(S_t, A_t, S_{t+1})] \coloneqq \frac{1}{n}\sum_{i=1}^n f(S_t^{(i)}, A_t^{(i)}, S_{t+1}^{(i)})$. 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:

$$Y_t(Q) \coloneqq R_t + \max_{a'}\left[Q(S_{t+1}, a')\right], \qquad Q_t(s,a) = \mathbb{E}[Y_t(Q_{t+1}) | S_t = s, A_t = a], \tag{10}$$

$$\hat{Q}_t \in \arg\min_{q_t \in \mathcal{Q}} \mathbb{E}_{n,t}[(Y_t(\hat{Q}_{t+1}) - q_t(S_t, A_t))^2]. \tag{11}$$

The Bayes-optimal predictor of Y_t is the true Q_t function, even though Y_t is a stochastic approximation of Q_t that replaces the expectation over the next-state transition with a stochastic sample thereof (realized from data). In this way, fitted-Q-iteration is *pseudo-outcome* regression, regressing onto a random variable whose conditional expectation is the target function, but is not equivalent to it under additive noise, as is the case with typical regression on observed outcomes. Pseudo-outcome regression has recently been used in causal inference [Kennedy, 2020, Semenova and Chernozhukov, 2021], and later in our robust procedure we are therefore able to use analogous arguments to obtain orthogonalized estimation. The procedure for FQE is exactly analogous, replacing the maximum over next-timestep actions with evaluation under the evaluation policy.

In our robust version of FQI, we instead approximate the robust Bellman operator from eq. (9). In particular, we will apply Proposition 2, but impose the constraints in eqs. (6) and (7) to arrive at the following optimization problem in terms of observable quantities:

Proposition 4. *Let Q be a real-valued function over states and actions, and define $Y_t(Q)$ as in Equation (11). Given Assumption 1 and Assumption 2, the robust $Q(s,a)$ function solves the following optimization problem:*

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

Next, in Section 4.1, we show that the optimization problem in Proposition 4 admits a closed form as a conditional expectation of observables. Then in Section 4.2, we incorporate this insight into an orthogonalized confounding-robust FQI algorithm with function approximation.

4.1 Closed-Form for the Robust Bellman Operator

Solving the optimization problem in Proposition 4 for each s, a pair isn't feasible for large state and action spaces. In this section, we use recent results to derive a *closed-form* expression for the minimum in Proposition 4 in order to derive a feasible algorithm leveraging function approximation. This is an application of the results in Rockafellar et al. [2000] and Dorn et al. [2021].

The closed-form state-action conditional solution to Proposition 4 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: $Z_{t,a}^{1-\tau} \coloneqq Z_t^{1-\tau}(Y_t(Q) \mid s,a), \alpha_t \coloneqq \alpha_t(S,A), \beta_t \coloneqq \beta_t(S,A)$. We can learn the conditional quantile functions by minimizing the *pinball loss* over a function class $\mathcal{Z}$:

$$Z_t^{1-\tau}(Y_t(Q) \mid S_t, A_t) \in \arg\min_{z \in \mathcal{Z}} \mathbb{E}[L_\tau(Y_t(Q), z(S_t,A_t))], \quad \text{where } L_\tau(y,\hat{y}) \coloneqq \begin{cases} (1-\tau)(\hat{y} - y), & \text{if } y < \hat{y} \\ \tau(y - \hat{y}), & \text{if } y \geq \hat{y} \end{cases}. \tag{12}$$

Proposition 5. *The solution to the minimization problem in Proposition 4 is:*

$$(\bar{\mathcal{T}}_t^* Q)(s,a) = \mathbb{E}_{obs} \left[\alpha_t Y_t(Q) + \frac{1-\alpha_t}{1-\tau} Y_t(Q) \mathbb{I}\left[Y_t(Q) \leq Z_{t,a}^{1-\tau} \right] \middle| S_t = s, A_t = a \right]. \tag{13}$$

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 Equation (7). 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 eq. (11).
    For $a \in \mathcal{A}$, where $A_t^{(i)} = a$, 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 eq. (14).
    For $a \in \mathcal{A}$, where $A_t^{(i)} = a$, 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

Proposition 5 suggests a simple two-stage procedure. First, estimate $Z_t^{1-\tau}$, and then estimate the conditional expectation in eq. (13) via regression using the estimated $Z_t^{1-\tau}$. We do so to develop robust policy evaluation and optimization algorithms in the next section. We first describe the basic method, its improvement via orthogonalization, and lastly sample splitting/cross-fitting.

Remark 1 (Estimating the Q-function, not the average policy value)**.** *This work focuses on policy optimization and evaluation via a closed-form solution of the robust Bellman operator and fitted-Q-iteration and targets estimation of the Q-function. The approach is different from robust generalizations of importance sampling in general; in Appendix C.3 we sketch an alternative analogous estimator of the policy value analogous to Jiang and Li [2016], Thomas et al. [2015], Namkoong et al. [2020] to illustrate the difference.*

4.2 Improving estimation: the orthogonalized pseudo-outcome

The two-stage procedure depends on the conditional quantile function $Z_t^{1-\tau}$, a *nuisance function* that must be estimated but is not our substantive target of interest. To avoid transferring biased first-stage estimation error of $Z_t^{1-\tau}$ to the Q-function, we introduce orthogonalization. Orthogonalized estimators remove the first-order dependence of estimating the target on the error in nuisance functions. An important literature from biostatistics and econometrics on Neyman-orthogonality (also called double/debiased machine learning, and related to semiparametric statistics) derives bias adjustments [Kennedy, 2022, Newey, 1994, Chernozhukov et al., 2018, Laan and Robins, 2003]. (See Appendix C for more). In particular, we apply an orthogonalization of Olma [2021] for what they call truncated conditional expectations, $m(\eta, x) = \frac{1}{1-\tau}\mathbb{E}[Y\mathbb{I}\left[Y \leq Z^{1-\tau}\right] \mid X = x]$. They show that

$$\tfrac{1}{1-\tau}\mathbb{E}[Y\mathbb{I}\left[Y \leq Z^{1-\tau}\right] - Z^{1-\tau}(\mathbb{I}\left[Y \leq Z^{1-\tau}\right] - (1-\tau)) \mid X]$$

is Neyman-orthogonal with respect to error in $Z^{1-\tau}$. Note that this comprises an additive, zero-mean adjustment to the original pseudo-outcome. We apply this orthogonalization to Equation (13) to obtain our regression target for robust FQE:

$$\tilde{Y}_t(Z, Q) \coloneqq \alpha_t Y_t(Q) + \tfrac{1-\alpha_t}{1-\tau}\Big(Y_t(Q)\mathbb{I}\left[Y_t(Q) \leq Z_t^{1-\tau}\right] - Z \cdot \{\mathbb{I}\left[Y_t(Q) \leq Z\right] - (1-\tau)\}\Big) \tag{14}$$

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:

$$\hat{\overline{Q}}_t \in \arg\min_{q_t} \mathbb{E}_{n,t}[(\tilde{Y}_t(\hat{Z}_t^{1-\tau}, \hat{\overline{Q}}_{t+1}) - q_t(S_t, A_t))^2]. \tag{15}$$

A large literature discusses methods for quantile regression [Koenker and Hallock, 2001, Meinshausen, 2006, Belloni and Chernozhukov, 2011], as well as conditional expected shortfall [Cai and Wang, 2008, Kato, 2012] and can guide the choice of function class for quantiles and $\overline{Q}$ appropriately.

We summarize the algorithm in Algorithm 1. In the appendix, we discuss a sample-splitting version in more detail; we describe the approach, which is standard, in the main text for brevity. Lastly, to ensure independent errors in nuisance estimation and the fitted-Q regression, for the theoretical results, we study a cross-time variant of the standard cross-fitting/sample-splitting scheme for orthogonalized estimation and machine learning. Interleaving between timesteps ensures downstream policy evaluation errors are independent of errors in nuisance evaluation at time t. Finally, we note that sample splitting can be avoided by posing Donsker-type assumptions on the function classes in the standard way. In the experiments (and algorithm description) in the interest of data-efficiency we do not data-split. Recent work of Chen et al. [2022] shows rigorously that sample-splitting may not be necessary under stability conditions; extending that analysis to this setting would be interesting future work.

4.3 Extension to continuous actions

Although the manuscript focuses on binary or categorical actions, the method can directly be extended to continuous action spaces, at the expense of sharpness results and interpretability of the robust set. Jesson et al. [2022] proposes a continuous-action sensitivity model which instead directly bounds the density ratio (rather than the odds ratio):

$$\frac{1}{\Lambda} \leq \frac{\pi_t^b(a \mid s)}{\pi_t^b(a \mid s, u)} \leq \Lambda \tag{16}$$

In the continuous setting, densities could be greater than 1, which would violate conditions on the odds ratio. One way to interpret this sensitivity parameter is via implications for the KL-divergence of nominal and complete propensity scores. We can readily apply this to our problem by changing the uncertainty set on W to that implied by the above. Namely, solve the same linear program of Proposition 4 but enforce that $W_t = \frac{\pi_t^b(a|s)}{\pi_t^b(a|x,u)}$ satisfy the constraints of eq. (16) rather than Assumption 2:

$$(\bar{\mathcal{T}}_t^* Q)(s,a) = \min_{W_t} \{ \mathbb{E}_{\text{obs}}[W_t Y_t(Q) | S_t = s, A_t = a] : \; \mathbb{E}_{\text{obs}}[W_t | S_t = s, A_t = a] = 1, \;\; \Lambda^{-1} \leq W_t \leq \Lambda^{-1}, \text{a.e.} \}.$$

That is, the characterization of Proposition 5 holds, replacing the (α_t, β_t) bounds arising from the MSM with (Λ^{-1}, Λ). The pointwise solution of the (s,a)-conditional optimization problem is structurally the same, i.e. a conditional quantile characterization at a different level. The only difference algorithmically is in the conditional quantile estimation; in the continuous action setting, we would appeal to function approximation and minimize the (orthogonalized) pinball loss of eq. (12) with the action as a covariate. In the infinite-data, nonparametric limit, this would be well-specified; in practice, there will be some additional approximation error. Given those conditional quantiles, the rest of the method, (orthogonalized, etc.) proceeds analogously as discussed previously.

5 Analysis and Guarantees

We first describe the estimation benefits we receive from orthogonalization before discussing analysis of robust fitted-Q-evaluation and iteration, and insights. (All proofs are in the appendix).

5.1 Estimation guarantees

We describe the orthogonalized estimation results, before the results about the full output of the robust fitted-Q-iteration. We also require some regularity conditions for estimation. We assume nonnegative bounded rewards throughout.

Assumption 3 (Estimation). *1. Nonnegative boundedness of outcomes:* $0 \leq R_t \leq B_R, \forall t$

We assume the transitions are continuously distributed, a common regularity condition for the analysis of quantiles.

Assumption 4 (Bounded conditional density). *Assume that* $P_t(s_{t+1} \mid s_t, a) < M_P, \forall t, s_t, s_{t+1}$ *a.s.*

We let $\hat{\mathbb{E}}_n$ indicate a function obtained by regression, on an appropriate data split independent of the nuisance estimation. Define

$$
\begin{aligned}
\hat{\overline{Q}}_t(s,a) &= \hat{\mathbb{E}}_n[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) \mid s,a] && \text{feasible regressed robust Q,}\\
\tilde{\overline{Q}}_t(s,a) &= \hat{\mathbb{E}}_n[\tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid s,a] && \text{oracle-nuisance regressed robust Q}\\
\overline{Q}_t(s,a) &= \mathbb{E}[\tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid s,a] && \text{oracle robust Q.}
\end{aligned}
$$

In the above, $\hat{\overline{Q}}_t(s,a) = \hat{\mathbb{E}}_n[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) \mid s,a]$ is the feasible *regressed* robust-Q-estimator with estimated nuisance $\hat{Z}$, while $\tilde{\overline{Q}}(s,a) = \hat{\mathbb{E}}_n[\tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid s,a]$ is the *regressed* robust-Q-estimator with *oracle* nuisance Z, and $\overline{Q}_t(s,a)$ is the true robust Q output at time t (relative to the future Q functions that are the output of the algorithm).

We assume the following regression stability assumption, which appears in Kennedy [2020]. It is a generalization of stochastic equicontinuity and is satisfied, for example, by nonparametric linear smoothers.

Assumption 5 (Regression stability). *Suppose $\mathcal{D}_1$ and $\mathcal{D}_2$ are independent training and test samples, respectively. Let: 1. $\widehat{f}(x) = \widehat{f}(x; \mathcal{D}_1)$ be an estimate of a function $f(x)$ using the training data $\mathcal{D}_1$, 2. $\widehat{b}(x) = \widehat{b}(x; \mathcal{D}_1) \equiv \mathbb{E}[\widehat{f}(x) - f(x) \mid \mathcal{D}_1, X = x]$ the conditional bias of the estimator $\widehat{f}$, 3. $\widehat{\mathbb{E}}_n[Y \mid X = x]$ denote a generic regression estimator that regresses outcomes on covariates in the test sample $\mathcal{D}_2$. Then the regression estimator $\widehat{\mathbb{E}}_n$ is defined as stable at $X = x$ (with respect to a distance metric d) if*

$$\frac{\widehat{\mathbb{E}}_n[\widehat{f}(x)|X=x] - \widehat{\mathbb{E}}_n[f(x)|X=x] - \widehat{\mathbb{E}}_n[\widehat{b}(x)|X=x]}{\sqrt{\mathbb{E}\left(\left[\widehat{\mathbb{E}}_n[f(x)|X=x] - \mathbb{E}[f(x)|X=x]\right]^2\right)}} \xrightarrow{p} 0$$

whenever $d(\widehat{f}, f) \xrightarrow{p} 0$.

Under these regularity conditions, we can show that the bias due to the first-stage estimation of the conditional quantiles is only quadratic in the estimation error of $\hat{Z}_t$.

Proposition 6 (CVaR estimation error)**.** *For $a \in \mathcal{A}, t \in [T]$, 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*

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

This implies we can maintain $o_p(n^{-\frac{1}{2}})$ consistent estimation of robust $\overline{Q}$ functions under weaker estimation error requirements on the conditional quantile functions Z.

Next, we describe key assumptions for convergence of fitted-Q-iteration, concentratability which restricts the distribution shift in the sequential offline data vs. optimized policies, and approximate Bellman completeness which assumes the closedness of the regression function class under the Bellman operator. Both these assumptions are standard requirements for fitted-Q-iteration, but certainly not innocuous; they do impose restrictions.

Assumption 6 (Concentratability)**.** *Given a policy π, let ρ_t^π denote the marginal distribution at time step t, starting from s_0 and following π, and μ_t denote the true marginal occupancy distribution under π^b. There exists a parameter C such that*

$$\sup_{(s,a,t)\in\mathcal{S}\times\mathcal{A}\times[T-1]} \frac{\mathrm{d}\rho_t^\pi}{\mathrm{d}\mu_t}(s,a) \leq C \quad \textit{for any policy } \pi.$$

Assumption 7 (Approximate Bellman completeness)**.** *There exists $\epsilon > 0$ such that, for all $t \in [T-1]$, where ϵ is at most on the order of $O_p(n^{-\frac{1}{2}})$,*

$$\sup_{q_{t+1}\in\mathcal{Q}_{t+1}} \inf_{q_t\in\mathcal{Q}_t} \|q_t - \overline{\mathcal{T}}_t^\star q_{t+1}\|_{\mu_t}^2 \leq \epsilon.$$

Concentratability is analogous to sequential overlap. It assumes a uniformly bounded density ratio between the true marginal occupancy distribution and those induced by arbitrary policies. Approximate Bellman completeness assumes that the function class $\mathcal{Q}$ is approximately closed under the robust Bellman operator. The requirement that ϵ is at most $O_p(n^{-\frac{1}{2}})$ is somewhat restrictive, but is also consistent with frameworks for local model misspecification that consider local asymptotics with $O_p(n^{-\frac{1}{2}})$ vanishing bias.

Although we ultimately seek an optimal policy, approaches based on fitted-Q-evaluation and iteration instead optimize the squared loss, which is related to the Bellman error that is a surrogate for value suboptimality.

Definition 2 (Bellman error)**.** *Under data distribution μ_t, define the Bellman error of function $q = (q_0, \ldots, q_{T-1})$ as: $\mathcal{E}(q) = \frac{1}{T}\sum_{t=0}^{T-1} \|q_t - \overline{\mathcal{T}}_t^* q_{t+1}\|_{\mu_t}$*

The next lemma, which appears as Duan et al. [2021, Lemma 3.2] (finite horizon), Xie and Jiang [2020, Thm. 2] (infinite horizon), justifies this approach by relating the Bellman error to the value suboptimality. Its proof follows immediately by considering the MDP given by the worst-case transition kernel that realizes the optimization in the definition of the robust Bellman operator and is omitted.

Lemma 1 (Bellman error to value suboptimality)**.** *Under Assumption 6, for any $q \in \mathcal{Q}$, we have that, for π the policy that is greedy with respect to q, $V_1^*(s_1) - V_1^\pi(s_1) \leq 2T\sqrt{C \cdot \mathcal{E}(q^\pi)}$.*

We will describe convergence results based on generic results for loss minimization over a function class of restricted complexity.

Definition 3 (Covering numbers, e.g. [van de Vaart and Wellner, 1996]). *Let $(\mathcal{F}, \|\cdot\|)$ be an arbitrary semimetric space. Then the covering number $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls of radius ϵ needed to cover $\mathcal{F}$.*

Definition 4 (Bracketing numbers). *Given two functions l and u, the bracket $[l, u]$ is the set of all functions f with $l \leq f \leq u$. An ϵ-bracket is a bracket $[l, u]$ with $\|u - l\| < \epsilon$. The bracketing number $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$.*

The covering and bracketing numbers for common function classes such as linear, polynomials, neural networks, etc. are well-established in standard references, e.g. Wainwright [2019], van de Vaart and Wellner [1996].

We assume either that the function class for $\mathcal{Q}, \mathcal{Z}$ is finite (but possibly exponentially large), or has well-behaved *covering* and *bracketing* numbers.

Assumption 8 (Finite function classes.). *The Q-function class $\mathcal{Q}$ and conditional quantile class $\mathcal{Z}$ are finite but can be exponentially large.*

Assumption 9 (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 1 (Fitted Q Iteration guarantee). *Suppose Assumptions 3, 4, 6 and 7 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 8 (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 9 (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

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

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

Finally, putting the above together with Lemma 1, our sample complexity bound states that the policy suboptimality is on the order of $O(n^{-\frac{1}{2}})$. Note that this analysis omits estimation error in π^b for simplicity.

The (log) covering number of many typical function classes is given in standard references [Wainwright, 2019, van de Vaart and Wellner, 1996]. Note that Lemma 9 gives that

$$N_{[]}\left(2\epsilon L, \mathcal{L}_{q(z'),z}, \|\cdot\|\right) \leq N(\epsilon, \mathcal{Q} \times \mathcal{Z}, \|\cdot\|) \leq N(\epsilon, \mathcal{Q}, \|\cdot\|) N(\epsilon, \mathcal{Z}, \|\cdot\|)$$

Therefore ensuring some $\epsilon = cn^{-\frac{1}{2}}$ approximation error (for some arbitrary constant c) can be achieved by fixing $\epsilon' = \frac{\epsilon}{2L}$; i.e. we require finer approximation.

Proof sketch. As appears elsewhere in the analysis of FQI [Duan et al., 2021], we may obtain the following standard decomposition:

$$\|\hat{\hat{Q}}_{t,\hat{Z}_t} - \overline{\mathcal{T}}^*_{t,\hat{Z}_t}\hat{\hat{Q}}_{t+1}\|^2_{\mu_t} = \mathbb{E}_\mu[\ell(\hat{\hat{Q}}_{t,\hat{Z}_t}, \hat{\hat{Q}}_{t+1}; \hat{Z}_t)] - \mathbb{E}_\mu[\ell(\overline{Q}^\dagger_{t,Z_t}, \hat{\hat{Q}}_{t+1}; Z_t)] + \|\overline{Q}^\dagger_{t,Z_t} - \overline{\mathcal{T}}^*_t\hat{\hat{Q}}_{t+1}\|^2_{\mu_t}$$

where $\overline{Q}^\dagger_{t,Z_t}$ is the oracle squared loss minimizer, relative to the $\hat{\hat{Q}}_{t+1}$ output from the algorithm. Assumption 7 (completeness) bounds the last term. Our analysis differs onwards with additional decomposition relative to estimated nuisances and applying orthogonality from Proposition 6.

Finally, we note that our analysis extends immediately to the infinite-horizon case, discussed in Appendix B.2 of the appendix due to space constraints. Crucially, the (s,a)-rectangular uncertainty set admits a stationary worst-case distribution [Iyengar, 2005].

5.2 Bias-variance tradeoff in selection of Λ

We can quantify the dependence of the sample complexity on constants related to problem structure. We consider an equivalent regression target which better illustrates this dependence.

Corollary 1. *Assume that the same function classes $\mathcal{Q}, \mathcal{Z}$ are used for every timestep, and they are VC-subgraph with dimensions v_q, v_z. Assume that $\epsilon_{\mathcal{Q},\mathcal{Z}} = 0$. Then, with r describing $L_r(M)$ norm under the discretization measure M, there exist absolute constants K, k such that*

$$\mathcal{E}(\hat{Q}) \leq K\{\log(v_q+v_z)+2(v_q+v_z)+r((v_q+v_z)-1)(T-1)\left(\log\left(2KB_r\Lambda(T-1)n/\epsilon\right)-1\right)\}n^{-1}+o_p(n^{-1}).$$

Note that the width of confidence bounds on the robust Q function in the horizon but logarithmically in Λ, which illustrates *robustness-variance-sharpness* tradeoffs. Namely, as we increase Λ, we estimate more extremal tail regions, which is more difficult. Sharper tail bounds on conditional expected shortfall estimation would also qualitatively yield similar insights.

5.3 Confounding with Infinite Data

While Theorem 1 analyses the difficulty of estimating the robust value function, here we analyze how the true robust value function differs from the nominal value function at the population-level for policy evaluation (not optimization). This gives a sense of how potentially conservative the method is, in case unconfoundedness held after all. We consider a simplified linear Gaussian setting.

Proposition 7. *Let $\mathcal{S} = \mathbb{R}$ and $\mathcal{A} = \{0, 1\}$. Define parameters $\theta_P, \theta_R, \sigma_P \in \mathbb{R}$. Suppose in the observational distribution that $S_{t+1}|S_t, A_t \sim \mathcal{N}(\theta_P S_t, \sigma_P)$, $R(s, a, s') = \theta_R s'$, $\pi^e_t(1|S_t) = 0.5$, and consider some π^b such that $\pi^b_t(A_t|S_t)$ does not vary with S_t. Finally, let $\beta_i := \theta_R \sum_{k=1}^{i} \theta_P^k$ and notice that the nominal, non-robust value functions are $V^{\pi^e}_{T-i}(s) = \beta_i s$ for $i \geq 1$. Then:*

$$|V^{\pi^e}_0(s) - \bar{V}^{\pi^e}_0(s)| \leq (16\theta_P)^{-1}(\textstyle\sum_{i=0}^{T-1}\beta_i)\sigma_P\log(\Lambda).$$

Note that the cost of robustness gets worse as the horizon T increases, depending on the value of θ_P. The parameter θ_P is the autoregressive coefficient for the state transitions — it controls how strongly last period's state impacts this period's state. In the language of linear systems, θ_P will determine whether or not the system is stable. Each of the stability regimes — stable, marginally stable, and unstable — results in different scaling with T for the cost of robustness. For $|\theta_P| < 1$, the term $(\sum_{i=0}^{T-1}\beta_i)/\theta_P$ is asymptotically linear in T; for $|\theta_P| = 1$, the term is quadratic in T; and for $|\theta_P| > 1$, the term scales asymptotically as θ_P^T. In other words, for *stable* systems, unobserved confounding can at worst induce bias that is linear in horizon, but for *unstable* systems, the bias could increase exponentially. In contrast, for the unconfounded problem, unstable systems are typically easier to estimate due to their better signal-to-noise ratio [Simchowitz et al., 2018]. While this example involves a scalar state for simplicity, we can straightforwardly generalize Proposition 7 to higher dimensions where the bias will depend on the spectrum of the transition matrix.

On the other hand, the scaling with the *degree* of confounding Λ is *independent of horizon*, and has a modest $\log(\Lambda)$ rate. This is surprising: it suggests that the horizon of the problem presents more of a challenge than the strength of confounding at each time step, and that T and Λ do not interact at the population level — at least in a simple linear-Gaussian setting. Characterizing exactly when the scaling with Λ is horizon-independent is a promising direction for future work.

6 Experiments

6.1 Simulation

In this section, we validate the performance of our estimator, including its scaling with the sensitivity parameter Λ and the importance of orthogonalization. Note that our goal is not to evaluate the utility of the marginal sensitivity model itself — we leave that to the existing empirical literature in medicine and social science. Instead, we demonstrate that our robust FQI procedure can successfully solve the MSM, validating our theoretical analysis. We perform simulation experiments in a mis-specified sparse linear setting with heteroskedastic conditional variance. Previous methods for sensitivity analysis in RL, Namkoong et al. [2020], Kallus and Zhou [2020a], Bruns-Smith [2021], *cannot* solve this continuous state setting with confounding at every time step. We use the following (marginal) data-generating process for the observational data:

$$\mathcal{S} \subset \mathbb{R}^d, \mathcal{A} = \{0, 1\}, S_0 \sim \mathcal{N}(0, 0.01), \qquad \pi^b(1|S_t) = 0.5,\ \forall S_t$$
$$P_{\text{obs}}(S_{t+1}|S_t, A_t) = \mathcal{N}(\theta_\mu S_t + \theta_A a, \max\{\theta_\sigma S_t + \sigma, 0\}), \qquad R(S_t, A_t, S_{t+1}) = \theta_R^T S_{t+1}$$

with parameters $\theta_\mu, \theta_\sigma \in \mathbb{R}^{d\times d}, \theta_R, \theta_A \in \mathbb{R}^d, \sigma \in \mathbb{R}$ chosen such that $AS_t + \sigma > 0$ with probability vanishingly close to 1. The number of features $d = 25$ and θ_μ and θ_σ are chosen to be column-wise sparse, with 5 and 20 non-zero columns respectively. We collect a dataset of size $n = 5000$ from a single trajectory. We then repeat this experiment in a higher-dimensional setting with $d = 100$ and $n = 600$ — the d/n ratio is 300 times worse.

We estimate $\bar{V}_1^*(s)$ for $T = 4$ and several different values of Λ, using both the orthogonalized and non-orthogonalized robust losses. For function approximation of the conditional mean and conditional quantile, we use Lasso regression. Note that while this is correctly specified in the non-robust setting, the CVaR is *non-linear* in the observed state due to the non-linear conditional standard deviation of $\theta_R^T S_{t+1}$, and therefore the Lasso is a misspecified model for the quantile and robust value functions. For details see Appendix E.3 in the Appendix.

We report the mean-squared error (MSE) of the value function estimate over 100 trials, alongside the average ℓ_2-norm parameter error and the percentage of the time a wrong action is taken. The MSE and percentage of mistakes compare the estimated value function/policy to an analytic ground truth and are evaluated on an independently drawn and identically distributed holdout sample of size $n = 200,000$ drawn from the initial state distribution. See the Appendix for details on the ground truth derivation.

Λ	Algorithm	MSE($\bar{V}_0^*$)	ℓ_2 Parameter Error	% wrong action
1	FQI	0.2927	2.506	0%
2	Non-Orthogonal	0.6916	3.458	5e-5%
	Orthogonal	0.4119	2.678	0%
5.25	Non-Orthogonal	10.87	7.263	0.39%
	Orthogonal	0.5552	3.110	0%
8.5	Non-Orthogonal	50.72	17.32	2.5%
	Orthogonal	0.7113	3.410	4e-5%
11.75	Non-Orthogonal	171.1	33.80	5.4%
	Orthogonal	1.336	3.666	6e-4%
15	Non-Orthogonal	432.9	55.86	8.2%
	Orthogonal	2.687	3.931	4e-3%

Table 1: Simulation results with $d = 25$ and $n = 5000$, 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.

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 2: 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.

The low-dimensional results in Table 1 illustrate two important phenomena. First, the MSE increases roughly quadratically with Λ. While in practice, we would like to certify robustness for higher levels of Λ, the estimated lower bounds become less reliable. Second, the non-orthogonal algorithm suffers from substantially worse mean-squared error and as a result selects a sub-optimal action more often, especially at high levels of Λ. Orthogonalization has a very large impact not just in theory, but in practice.

The results for the high-dimensional setting are in Table 2. In this setting, policy optimization is *substantially* harder — even the nominal policy estimate only picks the true optimal action 72% of the time. However, we still see almost identical behavior as in the low-dimensional setting when comparing the orthogonal and non-orthogonal estimators. Without orthogonalization, performance drops off dramatically as Λ increases, such that for $\Lambda = 15$, the policy is only slightly better than random choice. Our orthogonalized algorithm has MSE that decays more gracefully with Λ, and picks the correct action at essentially the same rate as the nominal algorithm, even as Λ increases.

Note that these simulation results validate our algorithm for *estimating* the worst-case value function and robust policy. They do not assess how quickly the *ground-truth* population robust value function decays with Λ. See Section 5.3 above for an initial discussion.

6.2 Complex real-world healthcare data

In the next computational experiments, we show how our method extends to more complex real-world healthcare data via a case study around the use of MIMIC-III data for off-policy evaluation of learned policies for the management of sepsis in the ICU with fluids and vasopressors [Larkin, 2023a]. Sepsis is an umbrella term for an extreme response to infection and is a leading cause of mortality, healthcare costs, and readmission. Still, the management of sepsis is complex and there remains substantial uncertainty about clinical guidelines [Evans et al., 2021]. Practitioners recommend dynamic changes in treatment, i.e. tracking the patient's state over time. For example, giving IV fluids is expected to be beneficial at the very beginning, but there are also expected risks from too much [Larkin, 2023b].The pioneering efforts in releasing the MIMIC-III database enabled the development of Markov decision process models via model-based approaches or offline reinforcement learning methods [Liu et al., 2020, Raghu et al., 2017, 2018, Lu et al., 2020, Rosenstrom et al., 2022]. However, a crucial challenge is *off-policy evaluation* for credible, data-driven estimates of the benefits of these learned policies, that are less vulnerable to model assumptions.

Crucial assumptions such as *unconfoundedness* are likely violated in this setting: treatment decisions probably included additional information not recorded in the database. (Indeed, the clinical literature certainly discusses other aspects of patient state and potential actions not included in the data). On the other hand, the comprehensive electronic health record (EHR) contains the most important factors in clinical decision-making such as patient vitals. So, our methods that develop *robust bounds* for off-policy evaluation of complex sequential policies can be applicable here, in highlighting the sensitivity of current learned policies to potential violations of sequential unconfoundedness. Since many research works used fitted-Q-iteration, we compare confounding-robust policies vs. naive policies for prescriptive insights.

We now describe the specific MDP data primitives. Following the data preprocessing of Killian et al. [2020] and cohort definition of Komorowski et al. [2018], the data covers an observation period of 72 hours past the onset of sepsis. Observed actions, administration of fluids or vaso-pressors, were categorized by volume and segmented into quantiles per each action type based on observational frequency. This leads to 25 possible discrete actions. Demographic and contextual features include age, gender, weight, ventilation and re-admission status. Other time-varying features include patient information such as blood pressure, heart rate, INR, various blood cell counts, respiratory rate, and different measures of oxygen levels (see Killian et al. [2020, Table 2] for exact description). The reward function takes on three values: $R = \{-1, 0, +1\}$ where -1 indicates patient death, $+1$ indicates leaving the hospital; and 0 for all other events.

6.2.1 Fitted-Q Iteration with Gradient Boosting

For this case study, we perform flexible non-parametric regression using gradient-boosted trees in place of the simple linear models in our earlier simulations [Friedman, 2001, Hastie et al., 2009]. Features include the full state vector and indicators for each action.

We begin with nominal (non-robust) estimation using standard fitted-Q iteration with gradient-boosted regression as our approximating function class. Implementing the robust estimator for MSM parameter Λ requires only a few simple modifications of nominal FQI with off-the-shelf tools. First, we estimate the behavior policy π^b using a gradient-boosted classifier. Then within the FQI loop, we estimate a conditional quantile model using gradient-boosted regression with the quantile loss, which is supported natively in the `scikit-learn` package. Finally, we use the estimated quantiles to compute the orthogonalized pseudooutcomes, and fit a model for the Q function with gradient-boosted regression. We compute the value functions and optimal policies for horizon up to $T = 11$.

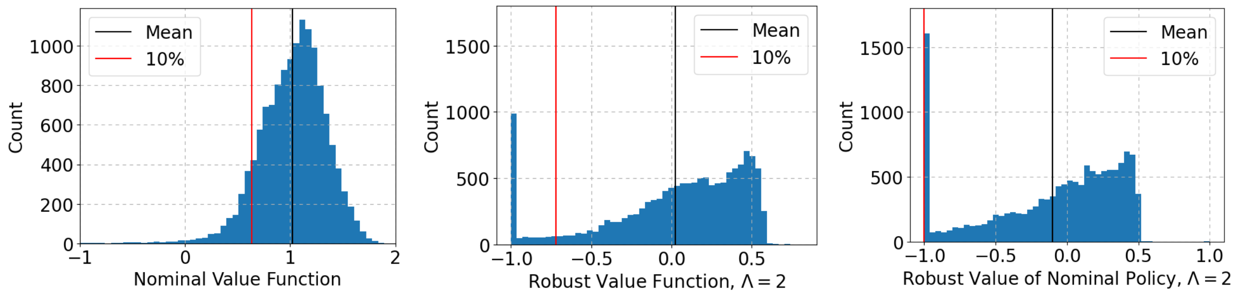
6.2.2 MIMIC Results

This case study is not meant to be a medical analysis, but concretely illustrates why caution is needed for interpreting offline RL applied to healthcare settings. In Figure 1a, we plot the distribution of the initial state value function, $V_0(s)$, with horizon $T = 11$ from non-robust FQI over the initial states in our dataset. The expected outcome under the nominal optimal policy is strongly positive for the majority of the population, including the 10% quantile.

By contrast, we plot the value function for the robust optimal value function (with $\Lambda = 2$) in Figure 1b. By construction, the robust value estimates are far more pessimistic. The *average* value of the robust optimal policy is still greater than zero, with a fairly substantial mass around $+0.5$. However, there is also a large negative tail with a strongly negative 10% quantile. We have truncated the plot at -1.0, which represents death, and notice that there are nearly 1000 starting states with value function ≤ -1.0. The more pessimistic outlook of the robust optimal value function represents the fact that some of the positive outcomes in the historical data could be due to spurious correlations with unobservables instead of a causal effect of the observed treatment.

We can also perform robust policy evaluation on the nominal optimal policy. We plot the corresponding value function over the initial states in Figure 1c. First, note that the expected robust value of the nominal optimal policy is actually negative. In other words, given only a modestly strong unobserved confounder ($\Lambda = 2$), it's possible that the nominal optimal policy *does more harm than good*. Furthermore, the number of initial states whose value is ≤ -1.0 has grown from about 1000 to about 1600, which now subsumes the 10% quantile. So under robust evaluation, not only does the nominal optimal policy have a slightly negative expected value for this distribution of patients, but it also substantially worsens the tail risk of death.

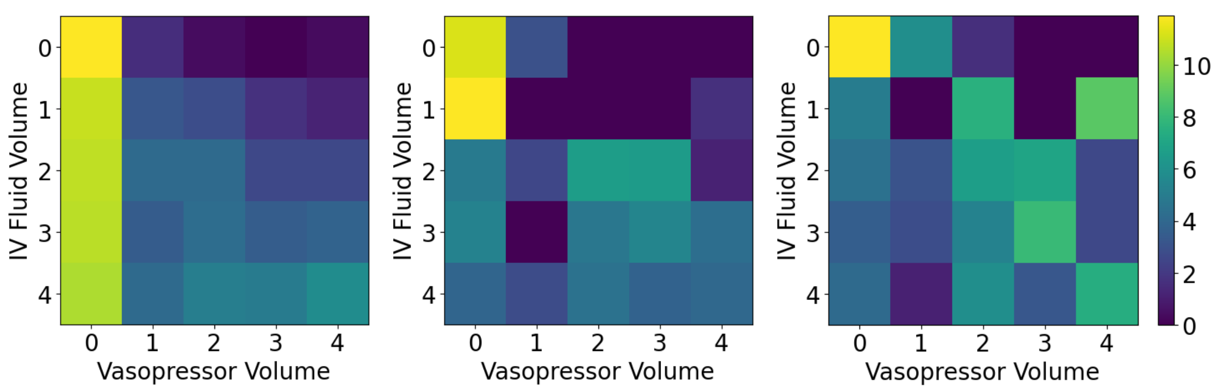
Beyond the value function, we also explore at a high-level how robustness changes the actions suggested by the optimal policy. In Figure 2, we compare the counts of actions taken in the historical data with the optimal actions from the nominal and robust policy. Figure 2a shows log counts of the historical actions, which includes a large number of patients with no treatment, many patients being treated with fluid but not vasopressors, and then a smaller number of patients receiving a variety of vasopressor intensities. The nominal optimal policy falls roughly the same pattern but made sharper; most patients are given either no treatment or the lowest level of IV fluid. Of the others, the majority are given a medium or large volume of both fluid and vasopressors. In contrast,



(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 $\Lambda = 2$. Each histogram includes a solid vertical line for the mean and the 10% quantile.

the robust optimal policy makes two key changes: there are more patients assigned to no treatment at all, but also more patients assigned to higher levels of vasopressors.



(a) Historical (b) Nominal Policy (c) Robust Policy, $\Lambda = 2$

Figure 2: Log of one plus counts of actions in the MIMIC-III dataset. The left panel plots the log counts of the actual actions observed, while the middle and right panels plot the log counts of the nominal and robust policy actions, respectively, given the observed states.

Finally, in Figure 3 we plot how the robust optimal actions change as the sensitivity parameter Λ is increased. At the far left, we have $\Lambda = 1$, which corresponds to the nominal policy, where a substantial fraction of patients are assigned to receiving only IV fluid. As Λ increases, the number of untreated increases dramatically, while the number treated with only fluid drops. At the same time, the number treated with both vasopressors and fluids increases by over ten times from $\Lambda = 1$ to $\Lambda = 2.5$. Note that we end the plot at $\Lambda = 2.5$. We find that higher values — even $\Lambda = 3$ — the robust value is mostly negative, with an large mass below -1.0. This reflects the fact that off-policy evaluation the MIMIC-III data is highly sensitive to unobserved variables.

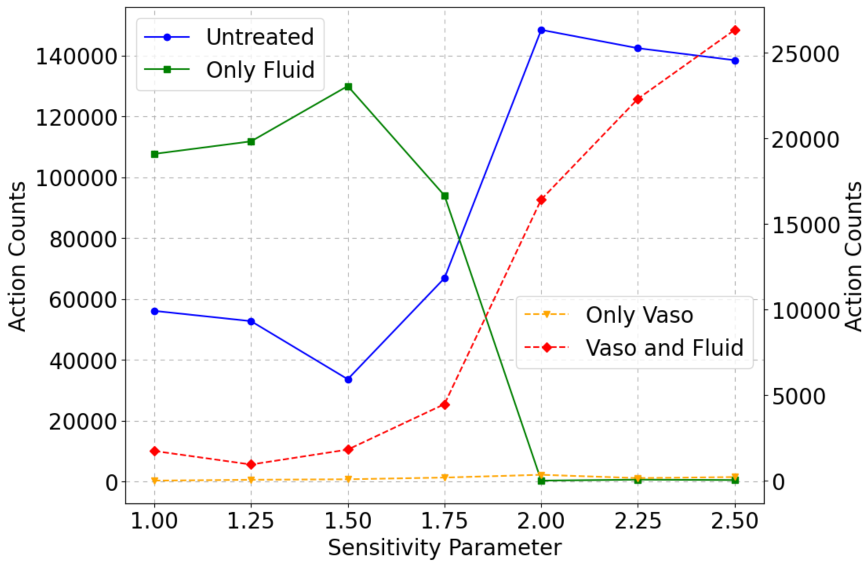


Figure 3: Counts of actions taken by the robust optimal policy over the states seen in the observed data as a function of the sensitivity parameter Λ. We combine the actions into four coarse groups: no treatment, only IV fluid, only vasopressors, and both fluid and vasopressors.

7 Extension: offline-online RL

In the previous sections, we discussed obtaining robust bounds from offline data for robust-optimal policy learning, via fitted-Q-iteration. However, even under the memoryless assumption of unobserved confounding, these bounds could be conservative due to the nature of compounding sequential uncertainties. Nonetheless, our robust approach and the structural memorylessness assumption makes it possible to leverage historical datasets to guide future randomized experimentation. We make this notion formal via *warmstarting* online learning procedures.

Importantly, we previously showed that the model of memoryless unobserved confounders results in a fully Markovian process over the observed states and actions. In the online setting, policies that do not depend on the unobserved confounder generate (unconfounded) Markov decision processes. We show how online RL algorithms based on optimism under uncertainty can improve performance by using information from valid robust bounds. This differs from modeling unobserved confounders in a generic POMDP, where standard RL algorithms do not apply even with online exploration.

In this section, we show how robust bounds can be used to warmstart a state-of-the-art reinforcement learning algorithm under linear function approximation, LSVI-UCB [Jin et al., 2020], a well-studied variant of least-squares value iteration (LSVI) [Bradtke and Barto, 1996, Osband et al., 2016] using linear function approximation. By contrast, naively (non-robustly) warmstarting LSVI-UCB by using confounded offline data severely degrades online performance.

Our work is most closely related to recent papers that warmstart reinforcement learning from offline data with unobserved confounding, although these have been restricted to tabular settings. Zhang and Bareinboim [2019] warm-start a variant of UCRL [Auer et al., 2008] for *tabular* dynamic treatment regimes with bounds from confounded data. Wang et al. [2021] does consider offline data with confounding and a similar warm-starting procedure. However, they also assume point-identifiability via backdoor adjustment or frontdoor adjustment. We will demonstrate that when this assumption fails, their procedure can have worse regret than not using the offline data at all. Other recent works, without unobserved confounders, study finer-grained hybrid offline-online RL [Xie et al., 2021b, Song et al., 2022]. [Tennenholtz et al., 2021] consider linear contextual bandits constrained by moment conditions from the offline data. Xu et al. [2023] studies restricted exploration for outperforming a conservative policy. We focus instead on demonstrating 1) how robust bounds from offline data can augment expensive online data and 2) how assuming memoryless unobserved confounders admits a marginal Markov decision process online counterpart, enabling warm-starting, unlike modeling unobserved confounders with POMDPs. We leave a full characterization for future work.

7.1 LSVI-UCB

We first introduce the basic setup of linear MDPs and LSVI-UCB [Jin et al., 2020]. As in our main problem setup, we assume both the online data and the offline data arise from the same underlying full-information Markov decision process. Under the memoryless unobserved confounders assumption, both the offline and online processes induce Markov decision processes over observables. We assume that the Q functions in the induced MDPs are *linear* and satisfy *completeness*. Let $\phi(s,a) : \mathcal{S} \times A \to \mathbb{R}^d$ be a feature map, and consider the function class $\mathcal{F}_{\text{lin}} \coloneqq \{f(s,a) = \langle \theta, \phi(s,a) \rangle : \theta \in \mathbb{R}^d\}$.

Assumption 10 (Linearity and Completeness)**.** *For any policy π^e that is only a function of the observed state, the Q function is linear, $Q_t^{\pi^e} \in \mathcal{F}_{lin}, \forall t$. Furthermore, for all $f \in \mathcal{F}$, we have the completeness condition:*

$$g(s,a) = \mathbb{E}_{S_{t+1}}[R_t + \max_{A'} f(S_{t+1}, A') | S_t = s, A_t = a] \in \mathcal{F}_{lin}, \forall t.$$

Under these assumptions, the online LSVI-UCB procedure of Jin et al. [2020] has total regret of order $\sqrt{T}$ with high probability. If the offline policy π^b is independent of the unobserved state u, then the online and offline MDPs are identical, and the setting reduces to one similar to Xie et al. [2021b]. However, if π^b does depend on the unobserved state, then the observed state transition probabilities will be different in the offline dataset. We will demonstrate that using our confounding-robust bounds, we can still use the offline dataset to warmstart LSVI-UCB, improving performance.

7.2 Warm-started LSVI-UCB

Here we outline the full algorithm for warm-starting LSVI-UCB presented in Algorithm 2. (Warm-starting other optimistic algorithms is essentially similar). The intuition is that the key step of LSVI-UCB, and other algorithms based on the principle of optimism under uncertainty, is planning according to the optimistic estimates of the value function, i.e. so that the estimated value function $\widehat{V}_t^n(s)$ satisfies that $\widehat{V}_t^n(s) \geq V_t^\star(s), \forall t, n, s, a$. This, in turn, bounds the per-episode regret by the difference between *optimistic* value function and true value function, $V_0^\star(s_0) - V_0^{\pi^n}(s_0) \leq \widehat{V}_0^n(s_0) - V_0^{\pi^n}(s_0)$. In the beginning, this difference is large due to sample uncertainty; but collecting more data over time shrinks the optimistic bonus and tends towards exploitation. Using the observational data, we can obtain *valid* robust bounds which can be used as a form of strong prior knowledge on the value function. That is, a basic idea is to truncate optimistic bounds by optimistic upper bounds over the confounded observational dataset. (Zhang and Bareinboim [2019] consider a similar approach but for tabular data). Truncating the optimistic bounds by prior knowledge 1) remains optimistic under valid bounds and 2) reduces the contribution of optimism to regret.

We now describe the basic algorithm in more detail. We run online LSVI-UCB, as in Jin et al. [2020] — each iteration we update our Q model and then collect a trajectory by taking actions that are optimal with respect to that Q model. The standard optimism bonus is $\xi\phi^T\Sigma_t^{-1}\phi$, where Σ_t is the sample Gram matrix and ξ is the width of the confidence interval; its value is derived theoretically but in practice it is often a hyperparameter. The key difference with standard LSVI-UCB is that at the start of each iteration, we run our robust FQE algorithm on the offline data to get robust upper bounds on the Q function for the current policy, $\hat{\overline{Q}}$. Note that since we want upper bounds instead of lower bounds, we compute the $\tau = \Lambda/(1+\Lambda)$ conditional quantile instead of the $1-\tau$ conditional quantile. Similarly, the $1-\tau$ terms in the pseudo-outcome in Equation (14) become τ, and $\mathbb{I}\left[Y_t(Q) \leq Z_t^{1-\tau}\right]$ becomes $\mathbb{I}\left[Y_t(Q) \geq Z_t^{1-\tau}\right]$.

Thus, each iteration we have two valid upper bounds on the Q function: the upper bound from the standard optimism bonus, and the upper bound from robust FQE on the offline data. For our warm-started LSVI-UCB, we choose whichever one is *smaller*. As a result, we retain the theoretical guarantees from optimism as proven in Jin et al. [2020], while possibly improving performance when $\hat{\overline{Q}}$ is sharper than the online upper confidence bound. For intuition, we can rewrite the robust value function as an upper confidence interval above the online point estimate to compare with the standard optimism bonus:

$$\begin{aligned} &\text{(online interval)} && \theta_t^\top \boldsymbol{\phi}(\cdot,\cdot) + \xi\left[\boldsymbol{\phi}(\cdot,\cdot)^\top \Sigma_t^{-1}\boldsymbol{\phi}(\cdot,\cdot)\right]^{1/2} \\ &\text{(robust offline interval)} && \theta_t^\top \boldsymbol{\phi}(\cdot,\cdot) + [\hat{\overline{Q}}_t(\cdot,\cdot) - \theta_t^\top \boldsymbol{\phi}(\cdot,\cdot)]_+ \end{aligned}$$

where $[\cdot]_+$ denotes the positive-part function. Recall that while our robust value function is a valid upper bound for the true value function, it need not be an upper bound for the online value function estimated from small samples. Thus, we simply set the offline bonus to 0 when $\hat{\overline{Q}}_t(\cdot,\cdot) < \theta_t^\top \boldsymbol{\phi}(\cdot,\cdot)$.

Finally, note that in practice, we can compute the robust *optimal* Q parameters once at the start using robust FQI, before the online procedure begins. By the definition of the robust optimal policy, the robust optimal Q function is always larger than the robust Q function of the online policy — thus using the robust optimal Q function is still a valid upper bound for the purposes of optimism. Formally, by saddlepoint properties, the policies evaluated by LSVI-UCB, $\hat{\pi}_k$, are feasible but suboptimal for the optimization problem that the robust Q function solves: since $(\bar{\pi}^*, \bar{P}_t^*) \in_\pi \inf_{\bar{P}_t \in \mathcal{P}_t} \mathbb{E}_{\bar{P}_t}\left[R_t + g\left(S_{t+1}, \pi_{t+1}^e\right) \mid s, a\right]$, we have that $\hat{\overline{Q}}_t \geq \hat{\overline{Q}}_t^{\hat{\pi}_k}$ (i.e. evaluating the latter at $\bar{P}_t^*$). This lets us perform offline robust FQI only once (instead of K times), which saves substantial computational cost at the expense of slightly looser upper bounds.

7.3 Simulation Experiments with Warm-starting

We provide preliminary experiments to demonstrate two key points. First, warmstarting LSVI-UCB from our valid robust bounds can result in substantial performance gains compared to the purely online algorithm. Second, naively warm-starting LSVI-UCB (without robustness) from confounded offline data performs much *worse* compared to the purely online algorithm.

For offline-online simulations, we consider a linear-gaussian MDP with an unobserved con-

Algorithm 2 Warm-Started LSVI-UCB

Estimate the marginal behavior policy, $\pi_t^b(a|s)$, in the offline data.
for episode $k = 1, \ldots, K$ **do**
  Initialize $\theta_T, \hat{\overline{Q}}_T = 0$
  for timestep $t = T-1, \ldots, 0$ **do**
    Estimate $\hat{\overline{Q}}_t$, robust Q function from observational dataset $\mathcal{D}_{obs}$, via robust policy eval for $\pi_t(\cdot) := \text{argmax}_a Q_{t+1}(\cdot, a)$, using the offline data as in Steps 4-6 of Algorithm 1
    $\Sigma_t \leftarrow \sum_{k'=1}^{k-1} \boldsymbol{\phi}\left(s_t^{k'}, a_t^{k'}\right) \boldsymbol{\phi}\left(s_t^{k'}, a_t^{k'}\right)^\top + \lambda \cdot \mathbf{I}$
    $\theta_t \leftarrow \Sigma_t^{-1} \sum_{k'=1}^{k-1} \boldsymbol{\phi}\left(s_t^{k'}, a_t^{k'}\right) \left[r_t^{k'} + \max_a Q_{t+1}\left(s_{t+1}^{k'}, a\right)\right]$
    $Q_t(\cdot,\cdot) \leftarrow \min\Big\{\theta_t^\top \boldsymbol{\phi}(\cdot,\cdot) + \xi\left[\boldsymbol{\phi}(\cdot,\cdot)^\top \Sigma_t^{-1} \boldsymbol{\phi}(\cdot,\cdot)\right]^{1/2},$
      $\max\{\theta_t^\top \boldsymbol{\phi}(\cdot,\cdot), \hat{\overline{Q}}_t(\cdot,\cdot)\},$
      $\mathrm{T}\Big\}$
  end for
  for step $t = 0, \ldots, T-1$ **do**
    Take action $a_t^k \leftarrow \pi_t^k(s_t^k) := \text{argmax}_{a \in \mathcal{A}} Q_h\left(s_t^k, a\right)$, and observe r_t^k and s_{t+1}^k
  end for
end for

founder U_t using the following parameterization:

$$\mathcal{S} \subset \mathbb{R}^8, \mathcal{A} = \{0,1,2,3\}, S_0 \sim \mathcal{N}(0, 0.1), \qquad \mathcal{U} = \{0,1,2,3\}, P_t(U_t|S_t) = 1/4$$
$$\pi^b(A_t|S_t, U_t) = 1/2 \text{ if } A_t = 3 - U_t,\ 1/6 \text{ otherwise} \implies \pi^b(A_t|S_t) = 1/4,$$
$$P_t(S_{t+1}|S_t, A_t, U_t) = \mathcal{N}(\theta_{\mu,s}S_t + \theta_{\mu,a}A_t + \theta_{\mu,u}U_t, \max\{\theta_{\sigma,s}S_t + \theta_{\sigma,a}A_t + 0.2, 0\}),$$
$$R_t = \mathcal{N}(\theta_{R,s}^T S_{t+1}, 10^{-8} + \mathbb{I}\left[U_t = 3\right]\mathbb{I}\left[A_t = 0\right]\sigma_R)$$

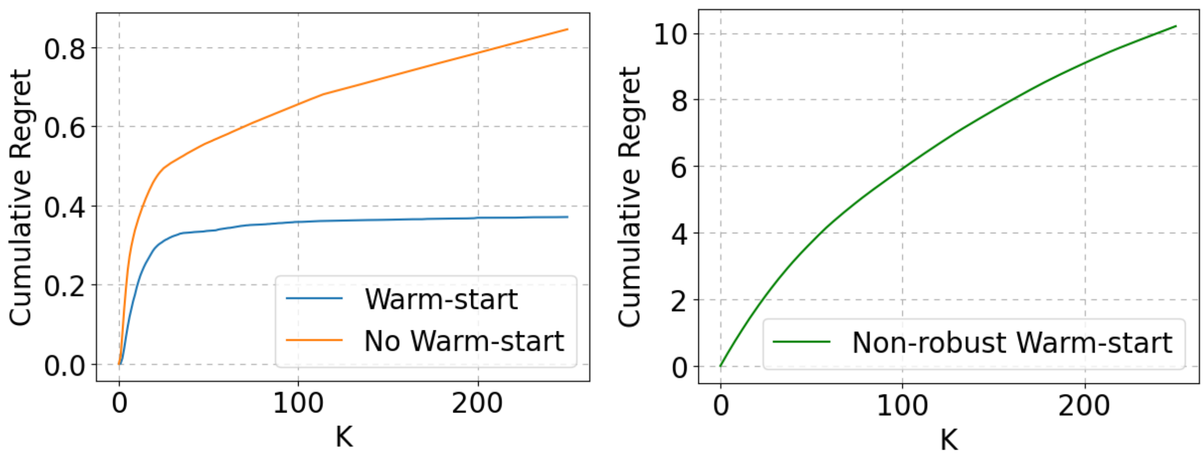
where the parameters $\theta_{\mu,s}, \theta_{\sigma,s} \in \mathbb{R}^{d \times d}$ and $\theta_{\mu,a}, \theta_{\mu,s}, \theta_{\sigma,a}, \theta_{R,s} \in \mathbb{R}^d$ are dense. Note that we've added some additional variability to the reward through the parameter $\sigma_R \in \mathbb{R}$; this is incorporated into our CVaR-based bounds without alteration because the variability is captured by the conditional quantile function. Finally, note that the smallest valid value for the MSM parameter is $\Lambda = 3$, as can be computed directly from $\pi^b(A_t|S_t, U_t)$ and $\pi^b(A_t|S_t)$.

Using this setup, we run the following three experiments: (1) standard LSVI-UCB without warm-starting, (2) warm-started LSVI-UCB using our robust bounds as in Algorithm 2, and (3) LSVI-UCB where we treat the offline data as if it were collected online and run the algorithm as usual. This third experiment will serve as our non-robust warm-starting benchmark - it is a simple (non-Bayesian) version of Algorithm 1 in Wang et al. [2021]. Note that if the data had in fact been generated online, then the upper confidence intervals for this non-robust warm-starting approach would be valid, and the LSVI-UCB performance guarantees would still hold. However, due to the unobserved confounders U_t, the resulting confidence intervals are not valid.

For all experiments, we use horizon $T = 4$, number of trajectories $K = 250$, and LSVI-UCB parameters $\xi = 0.07$ and $\lambda = 10^{-6}$. Note that ξ has to be set sufficiently large for standard LSVI-UCB to have a valid upper confidence interval, whereas our warm-starting bounds will result in a valid interval regardless of ξ, providing some additional robustness to hyperparameter tuning. See the Appendix for a discussion of results with different hyperparameters. We compare performance in terms of the cumulative regret:

$$\sum_{k=1}^{K}[V_0^*(s_0^k) - V_0^{\pi^k}(s_0^k)],$$

where V_t^* is the optimal value function.



(a)

(b)

Figure 4: Simulation results for online LSVI-UCB. Cumulative regret is an average over 200 trials. Panel (a) plots the cumulative regret of LSVI-UCB without warm-starting, and with robust warm-starting following Algorithm 2. Panel (b) plots the cumulative regret of LSVI-UCB where the offline data is naively treated as if it were collected online.

We plot the results in Figure 4. The y-axis displays the the cumulative regret averaged over 200 repeats of each algorithm. In Figure 4a, we compare the cumulative regret of LSVI-UCB without warm-starting and LSVI-UCB using our robust warm-starting algorithm. Our warm-started algorithm enjoys less than half the cumulative regret of standard LSVI-UCB after 250 online trajectories. In Figure 4b, we show results for naive warm-starting from offline data.The cumulative regret after 250 trajectories is >10 times higher than standard LSVI-UCB and >20 times higher than robust warm-starting. The offline data misleads non-robust warm-starting to confidently choose the wrong action, and it takes a substantial amount of online data collection to correct this.

7.4 Confidence intervals for unobserved confounding from finite observational datasets

For simplicity, so far we have described warm-starting with bounds obtained from a large observational dataset without finite-sample uncertainty in estimating bounds. We provide an asymptotic confidence interval under linear function approximation that readily extends our warmstarting approach to a finite observational study.

Let $\theta_t, \overline{\theta}_t$ be the parameter for the nominal and robust Q-function, respectively. We consider state-feature vectors, denoted as $\phi_{t,a} = \phi(S_t, a)$, i.e. they take a product form over actions for simplicity. We first require regularity conditions on the feature covariances.

Assumption 11 (Identification)**.** *Let* $\Sigma := \mathbb{E}[\phi(s,a)^\top \phi(s,a)]$ *denote population covariance matrix of state-action features. Assume that there exist* $0 < C_{\min} < C_{\max} < \infty$ *that do not depend on* d *s.t.* $C_{\min} \leqslant \min \operatorname{eig}(\Sigma) \leqslant \max \operatorname{eig}(\Sigma) \leqslant C_{\max}$ *for all* d*.*

Assumption 12 (Error of second moments)**.** *Let* $\epsilon = \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) - Q_t(s,a)$*. Assume lower and upper bounds on its second moments:* $0 < \underline{\sigma}^2 := \sup_{(s,a)\in(\mathcal{S}\times\mathcal{A})} \mathbb{E}\left[\epsilon^2 \mid s,a\right]$*, and* $\bar{\sigma}^2 := \sup_{(s,a)\in(\mathcal{S}\times\mathcal{A})} \mathbb{E}\left[\epsilon^2 \mid s,a\right] < 0$*.*

We show that orthogonality and cross-fitting yield asymptotic normality. Because of the backward recursive structure in estimation, our final asymptotic variance is that of estimation with generated regressors (i.e. the next-time-step Q function), which we analyze via the asymptotic variance of the generalized method of moments (GMM) [Newey and McFadden, 1994]. Let ζ denote the parameter for the linear conditional quantile. We overload notation and let $\tilde{Y}_{t,a}(\zeta_t^\top, \overline{\theta}_{t+1})$ denote the (a)-conditional pseudo-outcome with linear conditional quantile $Z_t = \zeta_t^\top \phi_t$ and robust Q function $\overline{Q}_t(s,a) = \overline{\theta}_t^\top \phi_t$, i.e. with a' the maximizing action or drawn with respect to the policy distribution.

Theorem 2 (Asymptotic normality for linear FQE)**.** *Under Assumptions 3, 6, 7, 11 and 12, the asymptotic covariance is defined via* $\bar{\theta}$ *satisfying the following moment equations: let*

$$g_{t,a}(\zeta^*, \bar{\theta}) = \left[\left\{\tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}_{t+1}) - \overline{\theta}_{t,a}^\top \phi_{t,a}\right\} \phi_{t,a}^\top\right] \mathbb{I}\left[A_t = a\right]/p(a), \tag{17}$$

then $\bar{\theta}$ *satisfies the stacked moment equation* $\{0 = \mathbb{E}[g_{t,a}(\zeta^*, \bar{\theta})]\}_{a\in\mathcal{A}, t=0,\ldots,T-1}$*.*

$$\sqrt{n}(\hat{\bar{\theta}} - \bar{\theta}^*) \xrightarrow{d} -\left(G^\top G\right)^{-1} G^\top \tilde{I}, \text{ where } \tilde{I} \sim N(0, I)$$

The matrix $G = \partial g(\zeta^*, \bar{\theta})/\partial \bar{\theta}$ *is an upper triangular matrix. The entries of* G *are as follows:*

$$\frac{\partial g_{t,a}(\zeta^*, \bar{\theta})}{\partial \bar{\theta}_{t,a}} = \mathbb{E}[\phi_{t,a}\phi_{t,a}^\top]$$

$$\frac{\partial g_{t,a}(\zeta^*, \bar{\theta})}{\partial \bar{\theta}_{t+1,a'}} = \mathbb{E}\left[\alpha_{t,a}(\phi_{t+1,a'}\phi_{t,a}^\top) + (1-\alpha_{t,a})(Z^{\phi}_{a',t,a}\phi_{t,a}^\top)\right]$$

$$\text{where } Z^{\phi}_{a_{t+1}}(S_t, a) = \mathbb{E}[\phi(S_{t+1}, a_{t+1}) \mid Y_{t+1} \leq \zeta_{t,a}^\top \phi_{t,a}, S_t, A_t = a].$$

Based on the asymptotic variance characterization, we can add an appropriate confidence interval to $\hat{\overline{Q}}$ in Step 7 of Algorithm 2 to maintain a high probability upper bound on the Q function.

8 Conclusion

We developed a robust fitted-Q-iteration algorithm under memory-less unobserved confounders, leveraging function approximation, conditional quantiles, and orthogonalization. Importantly, our algorithm can be implemented using only off-the-shelf tools by changing only a few lines of code of standard FQI, making it easily accessible to practitioners. We derived sample complexity guarantees, demonstrated the effectiveness of our algorithm and the benefits of orthogonality in simulation experiments, and then provided a case-study with complex real-world healthcare data. Finally, we showed how to use our robust bounds to warm-start online reinforcement learning, demonstrating substantial performance benefits, whereas naive use of the offline data for warm-starting can actually hurt performance. Interesting directions for future work include falsifiability-based analyses to draw on competing identification proposals, model-selection procedures for the conditional quantile and mean models, and a formal theoretical analysis of warm-starting with our robust bounds.

9 Acknowledgments

AZ acknowledges funding from the Foundations of Data Science Institute. This work was done in part while the author was visiting the Simons Institute for the Theory of Computing.

References

P. M. Aronow and D. K. Lee. Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika*, 100(1):235–240, 2013.

P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

A. Belloni and V. Chernozhukov. l1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

A. Bennett and N. Kallus. Policy evaluation with latent confounders via optimal balance. In *Advances in Neural Information Processing Systems*, pages 4827–4837, 2019.

A. Bennett, N. Kallus, L. Li, and A. Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pages 1999–2007. PMLR, 2021.

A. F. Bibaut, I. Malenica, N. Vlassis, and M. J. van der Laan. More efficient off-policy evaluation through regularized targeted learning. *arXiv preprint arXiv:1912.06292*, 2019.

M. Bonvini and E. H. Kennedy. Sensitivity analysis via the proportion of unmeasured confounding. *Journal of the American Statistical Association*, pages 1–11, 2021.

M. Bonvini, E. Kennedy, V. Ventura, and L. Wasserman. Sensitivity analysis for marginal structural models. *arXiv preprint arXiv:2210.04681*, 2022.

S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22:33–57, 1996.

D. A. Bruns-Smith. Model-free and model-based policy evaluation when causality is uncertain. In *International Conference on Machine Learning*, pages 1116–1126. PMLR, 2021.

Z. Cai and X. Wang. Nonparametric estimation of conditional var and expected shortfall. *Journal of Econometrics*, 147(1):120–130, 2008.

J. Chen and N. Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Q. Chen, V. Syrgkanis, and M. Austern. Debiased machine learning without sample-splitting for stable estimators. *arXiv preprint arXiv:2206.01825*, 2022.

S. Chen and B. Zhang. Estimating and improving dynamic treatment regimes with a time-varying instrumental variable. *arXiv preprint arXiv:2104.07822*, 2021.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

V. Chernozhukov, C. Cinelli, W. Newey, A. Sharma, and V. Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, 2022.

Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. In *Advances in Neural Information Processing Systems*, pages 1522–1530, 2015.

E. Delage and D. A. Iancu. Robust multistage decision making. In *The operations research revolution*, pages 20–46. INFORMS, 2015.

P. Ding, Y. Fang, D. Faries, S. Gruber, H. Lee, J.-Y. Lee, P. Mishra-Kalyani, M. Shan, M. van der Laan, S. Yang, et al. Sensitivity analysis for unmeasured confounding in medical product development and evaluation using real world evidence. *arXiv preprint arXiv:2307.07442*, 2023.

J. Dorn and K. Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, (just-accepted):1–28, 2022.

J. Dorn, K. Guo, and N. Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*, 2021.

Y. Duan, C. Jin, and Z. Li. Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pages 2892–2902. PMLR, 2021.

D. Ernst, G.-B. Stan, J. Goncalves, and L. 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.

L. Evans, A. Rhodes, W. Alhazzani, M. Antonelli, C. M. Coopersmith, C. French, F. R. Machado, L. Mcintyre, M. Ostermann, H. C. Prescott, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive care medicine*, 47(11): 1181–1247, 2021.

FDA. Aiml_samd_action_plan. `https://www.fda.gov/media/145022/download`, January 2021. (Accessed on 09/06/2023).

D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

J. Fu, M. Norouzi, O. Nachum, G. Tucker, Z. Wang, A. Novikov, M. Yang, M. R. Zhang, Y. Chen, A. Kumar, et al. Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*, 2021.

Z. Fu, Z. Qi, Z. Wang, Z. Yang, Y. Xu, and M. R. Kosorok. Offline reinforcement learning with instrumental variables in confounded markov decision processes. *arXiv preprint arXiv:2209.08666*, 2022.

O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, and L. A. Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

V. Goyal and J. Grand-Clement. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

S. Han. Optimal dynamic treatment regimes and partial welfare ordering. *Journal of American Statistical Association (just-accepted)*, 2022.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

J. Y. Hsu and D. S. Small. Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics*, 69(4):803–811, 2013.

E. Imani, E. Graves, and M. White. An off-policy policy gradient theorem using emphatic weightings. *Advances in Neural Information Processing Systems*, 31, 2018.

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

S. Jeong and H. Namkoong. Assessing external validity over worst-case subpopulations. *arXiv preprint arXiv:2007.02411*, 2020.

A. Jesson, A. Douglas, P. Manshausen, N. Meinshausen, P. Stier, Y. Gal, and U. Shalit. Scalable sensitivity and uncertainty analysis for causal-effect estimates of continuous-valued interventions. *arXiv preprint arXiv:2204.10022*, 2022.

N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

M. I. Jordan, Y. Wang, and A. Zhou. Empirical gateaux derivatives for causal inference. *arXiv preprint arXiv:2208.13701*, 2022.

N. Kallus and M. Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020a.

N. Kallus and M. Uehara. Statistically efficient off-policy policy gradients. In *International Conference on Machine Learning*, pages 5089–5100. PMLR, 2020b.

N. Kallus and A. Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *arXiv preprint arXiv:2002.04518*, 2020a.

N. Kallus and A. Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 2020b.

N. Kallus and A. Zhou. Stateful offline contextual policy evaluation and learning. In *International Conference on Artificial Intelligence and Statistics*, pages 11169–11194. PMLR, 2022.

N. Kallus, X. Mao, and A. Zhou. Interval estimation of individual-level causal effects under unobserved confounding. *arXiv preprint arXiv:1810.02894*, 2018.

K. Kato. Weighted nadaraya–watson estimation of conditional expected shortfall. *Journal of Financial Econometrics*, 10(2):265–291, 2012.

E. H. Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

T. W. Killian, H. Zhang, J. Subramanian, M. Fatemi, and M. Ghassemi. An empirical study of representation learning for reinforcement learning in healthcare. *arXiv preprint arXiv:2011.11235*, 2020.

R. Koenker and K. F. Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11): 1716–1720, 2018.

M. J. Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.

H. Larkin. Vasopressors or high-volume iv fluids both effective for sepsis. *JAMA*, 329(7):532–532, 2023a.

M. Larkin. Finding the Optimal Fluid Strategies for Sepsis — medscape.com. `https://www.medscape.com/viewarticle/993925?form=fpf#vp_2`, 2023b. [Accessed 27-08-2023].

H. Le, C. Voloshin, and Y. Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

G. Lewis and V. Syrgkanis. Double/debiased machine learning for dynamic treatment effects via g-estimation. *arXiv preprint arXiv:2002.07285*, 2020.

L. Liao, Z. Fu, Z. Yang, Y. Wang, M. Kolar, and Z. Wang. Instrumental variable value iteration for causal offline reinforcement learning. *arXiv preprint arXiv:2102.09907*, 2021.

Q. Liu, L. Li, Z. Tang, and D. Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366, 2018.

S. Liu, K. C. See, K. Y. Ngiam, L. A. Celi, X. Sun, and M. Feng. Reinforcement learning for clinical decision support in critical care: comprehensive review. *Journal of medical Internet research*, 22 (7):e18477, 2020.

E. A. Lobo, M. Ghavamzadeh, and M. Petrik. Soft-robust algorithms for batch reinforcement learning. *arXiv preprint arXiv:2011.14495*, 2020.

M. Lu, Z. Shahn, D. Sow, F. Doshi-Velez, and H. L. Li-wei. Is deep reinforcement learning ready for practical applications in healthcare? a sensitivity analysis of duel-ddqn for hemodynamic management in sepsis patients. In *AMIA Annual Symposium Proceedings*, volume 2020, page 773. American Medical Informatics Association, 2020.

X. Ma, Z. Liang, L. Xia, J. Zhang, J. Blanchet, M. Liu, Q. Zhao, and Z. Zhou. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

R. Miao, Z. Qi, and X. Zhang. Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. *arXiv preprint arXiv:2209.10064*, 2022.

L. W. Miratrix, S. Wager, and J. R. Zubizarreta. Shape-constrained partial identification of a population mean under unknown probabilities of sample selection. *Biometrika*, 105(1):103–114, 2018.

H. Namkoong, R. Keramati, S. Yadlowsky, and E. Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *arXiv preprint arXiv:2003.05623*, 2020.

W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

A. Nilim and L. El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

M. Norton, V. Khokhlov, and S. Uryasev. Calculating cvar and bpoe for common probability distributions with application to portfolio optimization and density estimation. *Annals of Operations Research*, 299(1):1281–1315, 2021.

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

L. Orellana, A. Rotnitzky, and J. M. Robins. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content. *The international journal of biostatistics*, 6(2), 2010.

I. Osband, B. Van Roy, and Z. Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016.

K. Panaganti, Z. Xu, D. Kalathil, and M. Ghavamzadeh. Robust reinforcement learning using offline data. *arXiv preprint arXiv:2208.05129*, 2022.

M. L. Puterman. *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, and M. Ghassemi. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.

A. Raghu, M. Komorowski, and S. Singh. Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602*, 2018.

P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

J. M. Robins, A. Rotnitzky, and D. 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.

R. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2: 21–42, 2000.

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

E. Rosenstrom, S. Meshkinfam, J. S. Ivy, S. H. Goodarzi, M. Capan, J. Huddleston, and S. Romero-Brufau. Optimizing the first response to sepsis: An electronic health record-based markov decision process model. *Decision Analysis*, 19(4):265–296, 2022.

S. Saghafian. Ambiguous dynamic treatment regimes: A reinforcement learning approach. *arXiv preprint arXiv:2112.04571*, 2021.

D. Scharfstein, A. McDermott, I. Diaz, M. Carone, N. Lunardon, and I. Turkoz. Global sensitivity analysis for repeated measures studies with informative drop-out: A semi-parametric approach. *Biometrics*, 74(1):207–219, 2018.

D. O. Scharfstein, R. Nabi, E. H. Kennedy, M.-Y. Huang, M. Bonvini, and M. Smid. Semiparametric sensitivity analysis: Unmeasured confounding in observational studies. *arXiv preprint arXiv:2104.08300*, 2021.

V. Semenova. Debiased machine learning of set-identified linear models. *arXiv preprint arXiv:1712.10024*, 2017.

V. Semenova. Debiased machine learning of set-identified linear models. *Journal of Econometrics*, 2023.

V. Semenova and V. Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

C. Shi, M. Uehara, J. Huang, and N. Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *International Conference on Machine Learning*, pages 20057–20094. PMLR, 2022a.

C. Shi, J. Zhu, S. Ye, S. Luo, H. Zhu, and R. Song. Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, pages 1–12, 2022b.

M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

R. Singh and V. Syrgkanis. Automatic debiased machine learning for dynamic treatment effects. *arXiv preprint arXiv:2203.13887*, 2022.

Y. Song, Y. Zhou, A. Sekhari, J. A. Bagnell, A. Krishnamurthy, and W. Sun. Hybrid rl: Using both offline and online data can make rl efficient. *arXiv preprint arXiv:2210.06718*, 2022.

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

Z. Tan. Model-assisted sensitivity analysis for treatment effects under unmeasured confounding via regularized calibrated estimation. *arXiv preprint arXiv:2209.11383*, 2022.

Z. Tang, Y. Feng, L. Li, D. Zhou, and Q. Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*, 2019.

G. Tennenholtz, S. Mannor, and U. Shalit. Off-policy evaluation in partially observable environments. *arXiv preprint arXiv:1909.03739*, 2019.

G. Tennenholtz, U. Shalit, S. Mannor, and Y. Efroni. Bandits with partially observable confounded data. In *Uncertainty in Artificial Intelligence*, pages 430–439. PMLR, 2021.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388, 2015.

M. Uehara, H. Kiyohara, A. Bennett, V. Chernozhukov, N. Jiang, N. Kallus, C. Shi, and W. Sun. Future-dependent value-based off-policy evaluation in pomdps. *arXiv preprint arXiv:2207.13081*, 2022.

A. van de Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Mathematics, 1996.

T. J. VanderWeele and P. Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274, 2017.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

L. Wang, Z. Yang, and Z. Wang. Provably efficient causal reinforcement learning with confounded observational data. *Advances in Neural Information Processing Systems*, 34:21164–21175, 2021.

S. Wang, N. Si, J. Blanchet, and Z. Zhou. A finite sample complexity bound for distributionally robust q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3370–3398. PMLR, 2023.

T. Xie and N. Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021a.

T. Xie, N. Jiang, H. Wang, C. Xiong, and Y. Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34: 27395–27407, 2021b.

W. Xu, Y. Ma, K. Xu, H. Bastani, and O. Bastani. Uniformly conservative exploration in reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 10856–10870. PMLR, 2023.

S. Yadlowsky, H. Namkoong, S. Basu, J. Duchi, and L. Tian. Bounds on the conditional and average treatment effect in the presence of unobserved confounders. *arXiv preprint arXiv:1808.09521*, 2018.

S. Yang and J. J. Lok. Sensitivity analysis for unmeasured confounding in coarse structural nested mean models. *Statistica Sinica*, 28(4):1703, 2018.

W. Yang, L. Zhang, and Z. Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6):3223–3248, 2022.

J. Zhang and E. Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, pages 13401–13411, 2019.

Q. Zhao, D. S. Small, and B. B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):735–761, 2019.

Y.-Q. Zhao, D. Zeng, E. B. Laber, and M. R. Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598, 2015.

Z. Zhou, Z. Zhou, Q. Bai, L. Qiu, J. Blanchet, and P. Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR, 2021.

Table of contents

- Appendix A includes proofs of results of the marginal MDP.
- Appendix B includes additional variants of the main algorithm (cross-fitting, infinite-horizon)
- Appendix C includes additional discussion.
- Appendix D includes proofs of analysis of robust FQI.
- Appendix E contains additional detail on computational experiments.
- Appendix F contains additional experiments in high-dimensions.

A Proofs for Section 3, Marginal MDP

Under Assumption 1, the setting with a policy π^e that only depends on the observed state is equivalent to a marginal MDP over the observed state alone:

Proposition 8 (Marginal MDP)**.** *Let χ^{marg} be the marginal distribution of χ over the observed state. Given Assumption 1, there exists $P_t^{marg} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ such that for any policies π^e and $\pi^{e'}$ that do not depend on U_t and for all s, a, t:*

$$P_t^{marg}(s, a) = \mathbb{P}_{\pi^e}(S_{t+1}|S_t = s, A_t = a) = \mathbb{P}_{\pi^{e'}}(S_{t+1}|S_t = s, A_t = a).$$

Furthermore, we can define a new MDP, $(\mathcal{S}, \mathcal{A}, R, T^{marg}, \chi^{marg}, H)$, with probabilities under policy π^e denoted $\mathbb{P}_{\pi^e}^{marg}$ such that $\forall \pi^e$,

$$\mathbb{P}_{\pi^e}^{marg}(S_0, A_0, ...S_H, A_H) = \mathbb{P}_{\pi^e}(S_0, A_0, ...S_H, A_H).$$

For any t, let $H_t = \{S_j, A_j : j \leq t\}$ be the history of the *observed* state and actions up to time t. In the rest of this section, we will use shorthands like $\mathbb{P}_\pi(s_{t+1}|s_t, a_t, h_{t-1}) := \mathbb{P}_\pi(S_{t+1}|S_t = s_t, A_t = a_t, H_{t-1} = h_{t-1})$ whenever clear from the text.

We will prove a more general version of Proposition 8, under the assumption that the observed state transition probabilities are Markovian. Markovian observed state transitions follow are guaranteed under Assumption 1:

Lemma 2 (Observed State Markov Property)**.** *Given Assumption 1, for all t, s, a, h,*

$$\mathbb{P}_\pi(s_{t+1}|s_t, a_t, h_{t-1}) = \mathbb{P}_\pi(s_{t+1}|s_t, a_t)$$
$$\mathbb{P}_\pi(a_t|s_t, h_{t-1}) = \mathbb{P}_\pi(a_t|s_t)$$

As alluded to in the main text, Lemma 2 is an implication of Assumption 1 that can be tested from observed states and actions alone. Although violations of Lemma 2 (i.e. non-Markovianness of observational transitions) rule out Assumption 1, the inverse is not true: a Markovian observational confounded distribution neither implies Assumption 1 nor the absence of unobserved confounders in general. We will use the following two Lemmas to prove Lemma 2:

Lemma 3. *Given Assumption 1, for any s_t, a_t, h_{t-1}, π:*

$$\mathbb{P}_\pi(a_t|s_t, h_{t-1}) = \mathbb{P}_\pi(a_t|s_t).$$

Proof. Proof of Lemma 3

$$\begin{aligned}\mathbb{P}_\pi(a_t|s_t,h_{t-1}) &= \int_{\mathcal{U}} \mathbb{P}_\pi(a_t|s_t,u_t,h_{t-1})\mathbb{P}_\pi(u_t|s_t,h_{t-1})du\\ &= \int_{\mathcal{U}} \mathbb{P}_\pi(a_t|s_t,u_t)P_t^u(u_t|s_t)du\\ &= \mathbb{P}_\pi(a_t|s_t),\end{aligned}$$

where we used that $U_t \sim P_u(S_t)$ independently of the past in the second line. □

Lemma 4. *Given Assumption 1, for any s_t, a_t, h_{t-1}, π:*

$$\mathbb{P}_\pi(u_t|s-t,a_t,h_{t-1}) = \mathbb{P}_\pi(u_t|s-t,a_t) = \frac{\pi_t(a_t|s_t,u_t)}{\pi_t(a_t|s_t)}P_t^u(u_t|s_t).$$

Proof. Proof of Lemma 4

$$\begin{aligned}\mathbb{P}_\pi(u_t|s_t,a_t,h_{t-1}) &= \frac{\mathbb{P}_\pi(a_t|s_t,u_t,h_{t-1})\mathbb{P}_\pi(u_t|s_t,h_{t-1})}{\mathbb{P}_\pi(a_t|s_t)}\\ &= \frac{\mathbb{P}_\pi(a_t|s_t,u_t)\mathbb{P}_\pi(u_t|s_t)}{\mathbb{P}_\pi(a_t|s_t)}\\ &= \frac{\pi_t(a_t|s_t,u_t)}{\pi_t(a_t|s_t)}P_t^u(u_t|s_t),\end{aligned}$$

where the second line follows because by the definition of an MDP, $\mathbb{P}_\pi(a_t|s_t,u_t,h_{t-1})$, and $U_t \sim P_u(S_t)$ is independent of the past. □

Proof. Proof of Lemma 2 We have:

$$\begin{aligned}\mathbb{P}_\pi(s_{t+1}|s_t,a_t,h_{t-1}) &= \int_{\mathcal{U}} \mathbb{P}_\pi(s_{t+1}|s_t,u_t,a_t,h_{t-1})\mathbb{P}_\pi(u_t|s_t,a_t,h_{t-1})du\\ &= \int_{\mathcal{U}} \mathbb{P}_\pi(s_{t+1}|s_t,u_t,a_t)\mathbb{P}_\pi(u_t|s_t,a_t)du\\ &= \mathbb{P}_\pi(s_{t+1}|s_t,a_t),\end{aligned}$$

where the second line applies Lemma 4. The proof then follows together with Lemma 3. □

We now prove the more general version of Proposition 8:

Proposition 9 (Marginal MDP, General)**.** *Let χ^{marg} be the marginal distribution of χ over the observed state. If for all s, a, h, t,:*

$$\mathbb{P}_\pi(S_{t+1}|S_t = s, A_t = a, H_{t-1} = h) = \mathbb{P}_\pi(S_{t+1}|S_t = s, A_t = a), \tag{18}$$

then there exists $P_t^{marg} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ such that for any policies π^e and $\pi^{e'}$ that do not depend on U_t and for all s, a, t:

$$P_t^{marg}(s,a) = \mathbb{P}_{\pi^e}(S_{t+1}|S_t = s, A_t = a) = \mathbb{P}_{\pi^{e\prime}}(S_{t+1}|S_t = s, A_t = a).$$

Furthermore, we can define a new MDP, $(\mathcal{S}, \mathcal{A}, R, T^{marg}, \chi^{marg}, H)$, with probabilities under policy π^e denoted $\mathbb{P}_{\pi^e}^{marg}$ such that

$$\mathbb{P}_{\pi^e}^{marg}(S_0, A_0, ...S_H, A_H) = \mathbb{P}_{\pi^e}(S_0, A_0, ...S_H, A_H).$$

The proof uses the following two lemmas:

Lemma 5 (Conditional Mean Independence with Respect to Transitions)**.** *Given Equation* (18)*,*

$$\int_{\mathcal{U}} \mathbb{P}_\pi(u_t|s_t, h_{t-1})\mathbb{P}_\pi(s_{t+1}|s_t, a_t, u_t)du = \int_{\mathcal{U}} \mathbb{P}_\pi(u_t|s_t)\mathbb{P}_\pi(s_{t+1}|s_t, a_t, u_t)du.$$

Proof. Proof of Lemma 5 Note that the full-information state transitions are Markovian by the definition of an MDP:

$$\mathbb{P}_\pi(s_{t+1}|s_t, a_t, u_t, h_{t-1}) = \mathbb{P}_\pi(s_{t+1}|s_t, a_t, u_t).$$

The lemma then follows by applying the tower property to both sides of Assumption 1. □

Lemma 6. *Given Equation* (18)*, for any two π^e and $\pi^{e'}$ which do not depend on U, $\forall s, a$, and t:*

$$\mathbb{P}_{\pi^e}(s_{t+1}|s_t, a_t) = \mathbb{P}_{\pi^{e'}}(s_{t+1}|s_t, a_t).$$

Proof. Proof of Lemma 6 The proof proceeds by mutual induction on the statement above and the following statement:

$$\mathbb{P}_{\pi^e}(u_t|s_t, h_{t-1}) = P\mathbb{P}\pi^{e'}(u_t|s_t, h_{t-1}).$$

We will consider π^e and demonstrate the equality with $\pi^{e'}$ by showing that the relevant qualities do not depend on π^e. First, consider $t = 0$. From the definition of the initial state distribution,

$$\mathbb{P}_{\pi^e}(u_0|s_0) = \chi(u_0|s_0).$$

which holds for all π^e.

From the definition of the MDP, $\mathbb{P}_\pi(s_{t+1}|s_t, u_t, a_t) = P_t(s_{t+1}|s_t, u_t, a_t)$ for any π. Then we have:

$$\begin{aligned}
\mathbb{P}_{\pi^e}(s_1|s_0, a_0) &= \int_{\mathcal{U}} \mathbb{P}_{\pi^e}(u_0|s_0, a_0)\mathbb{P}_{\pi^e}(s_1|s_0, a_0, u_0)du \\
&= \int_{\mathcal{U}} \mathbb{P}_{\pi^e}(u_0|s_0, a_0)P_0(s_1|s_0, a_0, u_0)du \\
&= \int_{\mathcal{U}} \mathbb{P}_{\pi^e}(u_0|s_0)P_0(s_1|s_0, a_0, u_0)du \\
&= \int_{\mathcal{U}} \chi(u_0|s_0)P_0(s_0, a_0, u_0)du,
\end{aligned}$$

where the third equality uses the fact that π^e does not depend on U. This equality also holds for all π^e and so we have proven the base case.

Now we consider a general t:

$$\begin{aligned}
\mathbb{P}_{\pi^e}(u_t|s_t, h_{t-1}) &= \int_{\mathcal{U}} \mathbb{P}_{\pi^e}(u_{t-1}|s_t, h_{t-1})P_{\pi^e}(u_t|s_t, h_{t-1}, u_{t-1})du \\
&= \int_{\mathcal{U}} \mathbb{P}_{\pi^e}(u_{t-1}|s_{t-1}, h_{t-2})\frac{\mathbb{P}_{\pi^e}(s_t = s, u_t|s_{t-1}, u_{t-1}, a_{t-1})}{\mathbb{P}_{\pi^e}(s_t|s_{t-1}, a_{t-1})}du,
\end{aligned}$$

where the second equality follows from applying Bayes rule to both probabilities in the second line. By the inductive hypothesis, $\mathbb{P}_{\pi^e}(u_{t-1}|s_{t-1}, h_{t-2})$ does not depend on π^e. The transition probabilities $\mathbb{P}_{\pi^e}(s_t, u_t | s_{t-1}, u_{t-1}, a_{t-1})$ do not depend on π^e. And by the inductive hypothesis, $\mathbb{P}_{\pi^e}(s_t | s_{t-1}, a_{t-1})$ does not depend on π^e. Therefore, $\mathbb{P}_{\pi^e}(u_t | s_t, h_{t-1})$ does not depend on π^e.

Finally,

$$\begin{aligned}
\mathbb{P}_{\pi^e}(s_{t+1}|s_t, a_t) &= \int_{\mathcal{U}} P_{\pi^e}(u_t|s_t, a_t) P_{\pi^e}(s_{t+1}|s_t, u_t, a_t) du \\
&= \int_{\mathcal{U}} \mathbb{P}_{\pi^e}(u_t|s_t, a_t) P_t(s_{t+1}|s_t, u_t, a_t) du \\
&= \int_{\mathcal{U}} \mathbb{P}_{\pi^e}(u_t|s_t) P_t(s_{t+1}|s_t, u_t, a_t) du \\
&= \int_{\mathcal{U}} \mathbb{P}_{\pi^e}(u_t|s_t, h_{t-1}) P_t(s_{t+1}|s_t, u_t, a_t) du
\end{aligned}$$

where the third equality follows from the fact that π^e does not depend on U and the fourth equality follows from Lemma 1. We have already shown that $P_t(s_{t+1}|_{\pi^e}(u_t|s_t, h_{t-1})$ does not depend on π^e which concludes the proof.

□

Proof. Proof of Proposition 9 Define $P_t^{\text{marg}} = \mathbb{P}_{\pi^e}(s_{t+1}|s_t, a_t)$, which by Lemma 6 is the same for any π^e. From the conditional independence structure of the original MDP together with Assumption 1, we have

$$\begin{aligned}
\mathbb{P}_{\pi^e}(S_0, A_0, ..., S_H, A_H) &= \mathbb{P}_{\pi^e}(S_0)\mathbb{P}_{\pi^e}(A_0|S_0) \prod_{t=1}^{H} \mathbb{P}_{\pi^e}(A_t|S_t)\mathbb{P}_{\pi^e}(S_t|S_{t-1}, A_{t-1}) \\
&= \chi^M(S_0)\pi^e(A_0|S_0) \prod_{t=1}^{H} \pi^e(A_t|S_t) P_t^{\text{marg}}(S_t|S_{t-1}, A_{t-1}) \\
&= \mathbb{P}_{\pi^e}^M(S_0, A_0, ...S_H, A_H).
\end{aligned}$$

□

Proof. Proof of Proposition 8 The result follows from Lemma 2 and Proposition 9. □

Note that we do not need to make Assumption 1 in order to construct the marginal MDP. Instead we can use Equation (18) directly. However, the next proof is substantially simpler with Assumption 1:

Proof. Proof of Proposition 2

$$\begin{aligned}
\mathbb{E}_{S'\sim T(s,a)}[f(s,a,S')] &= \mathbb{E}_{\pi^e}[f(S,A,S')|S=s,A=a] \\
&= \int_{\mathcal{S}} f(s,a,s')\mathbb{P}_{\pi^e}(S'=s'|S=s,A=a)ds' \\
&= \int_{\mathcal{S}} f(s,a,s')\left(\int_{\mathcal{U}} \mathbb{P}_{\pi^e}(U=u|S=s,A=a)\mathbb{P}_{\pi^e}(S'=s'|S=s,A=a,U=u)du\right)ds' \\
&= \int_{\mathcal{S}} f(s,a,s')\left(\int_{\mathcal{U}} \mathbb{P}_{\pi^e}(U=u|S=s)P_t(s'|s,a,u)du\right)ds' \\
&= \int_{\mathcal{S}} f(s,a,s')\left(\int_{\mathcal{U}} P_t^u(U=u|S=s)P_t(s'|s,a,u)du\right)ds' \\
&= \int_{\mathcal{S}} f(s,a,s')\left(\int_{\mathcal{U}} \frac{\pi^b(a|s)}{\pi^b(a|s,u)}\mathbb{P}_{\pi^b}(U_t=u|S_t=s,A_t=a)P_t(s'|s,a,u)du\right)ds' \\
&= \mathbb{E}_{\pi^b}\left[\frac{\pi^b(A|S)}{\pi^b(A|S,U)}f(S,A,S')\middle|S=s,A=a\right].
\end{aligned}$$

where the second to last equality follows by Lemma 4. □

We conjecture that the same result would hold under Equation (18), but it would require showing that

$$\int_{\mathcal{U}} \mathbb{P}_{\pi^e}(u|s)P_t(s'|s,a,u)du = \int_{\mathcal{U}} \mathbb{P}_{\pi^b}(u|s)P_t(s'|s,a,u)du$$

Note that: $\mathbb{P}_\pi(u|s) = \mathbb{P}_\pi(u|s,h)$ when under the integral with the transitions. So we need to use the fact that this is history-independent.

Proof. Proof of Proposition 4 The result follows by applying Corollary 4 of Dorn and Guo [2022] to Proposition 2. □

A.1 Realizability Counterexample

We'll consider a highly simplified empirical distribution with only a single state. We'll drop all dependences on S and t for simplicity. The possible outcomes Y lie in a discrete set and each have equal probability. We have three actions, the first with 4 data points, the second with 8 data points, and the last with 12 data points:

$$\begin{aligned}
&N = 24 \\
&P(A=0) = 4/24, P(A=1) = 8/24, P(A=2) = 12/24
\end{aligned}$$

Let the outcomes for the four $A=0$ datapoints be $\{Y_i = i : i \text{ from 1 to 4}\}$. Similarly Y_j and Y_k for $A=1$ and $A=2$ respectively. Then:

$$P(Y_i|A=0) = 1/4, P(Y_j|A=1) = 1/8, P(Y_k|A=2) = 1/12$$

Let $\Lambda = 3$, so that $1 - \tau = 1/4$. Denote the relevant lower bounds on the weights as $\alpha(A) = P(A) + \frac{1}{\Lambda}(1 - P(A))$ and $\beta(A) = P(A) + \Lambda(1 - P(A))$. Then from the Dorn and Guo result, we have unique weights that achieve the infimum over the MSM ambiguity set:

$$\begin{aligned}
&\text{For } A = 0, w = \{\beta(0), \alpha(0), \alpha(0), \alpha(0)\},\\
&\text{For } A = 1, w = \{\beta(1), \beta(1), \alpha(1), \alpha(1), \alpha(1), \alpha(1), \alpha(1), \alpha(1)\},\\
&\text{For } A = 2, w = \{\beta(2), \beta(2), \beta(2), \alpha(2), \alpha(2), \alpha(2), \alpha(2), \alpha(2), \alpha(2), \alpha(2), \alpha(2), \alpha(2)\}
\end{aligned}$$

Consider the first weight for $A = 0$, $w = \beta(0)$. We know that there exists some arbitrary u such that $P(A = 0)/P(A = 0|U = u) = \beta(0)$. Bayes rule then implies that:

$$P(U = u) = P(U = u|A = 0)\beta(0)$$

Then we have:

$$\begin{aligned}
&P(U = u) = P(U = u|A = 0)\beta(0) = \sum_{a} p(A = a)p(U = u|A = a)\\
\implies &P(U = u|A = 0)\beta(0) - P(U = a|A = 0)P(A = 0) = \sum_{a \neq 0} P(A = a)P(U = u|A = a)
\end{aligned}$$

and since $\beta(0) > p(A = 0)$, the probability of u occurring in the other actions must be non-zero. We therefore know that $P(A = 1|U = u) \in \{P(A = 1)/\alpha(A = 1), P(A = 1)/\beta(a = 1)\}$ and similarly for $P(A = 2|U = u)$. But there does not exist any choice such that $\sum_a P(A = a|U = u) = 1$ given our choices of Λ and $P(A)$.

B Algorithm Variants

B.1 With cross-fitting

Algorithm 3 Confounding-Robust Fitted-Q-Iteration

1: Initialize $\hat{\overline{Q}}_T = 0$. Obtain index sets of cross-fitted folds, $\{\mathcal{I}_{k(i,t)}\}_{i\in[K],t\in[T]}$
2: **for** $t = T-1,\ldots,1$ **do**
3: Using data $\{\left(S_t^i, A_t^i, R_t^i, S_{t+1}^i\right) : k(i,t) = k'\}$:
  Estimate the marginal behavior policy $\pi_t^b(a|s)$ and evaluate bounds $\alpha_t(s_t,a_t), \beta_t(s_t,a_t)$ as in Equation (7).
  Compute nominal outcomes $\{Y_t^{(i)}(\hat{\overline{Q}}_{t+1}^{-k'})\}_{i=1}^n$ as in eq. (11).
  For all $a \in \mathcal{A}$, fit $\hat{Z}_t^{1-\tau,k'}(s,a)$ the $(1-\tau)$th conditional quantile of the outcomes $Y_t^{(i)}$.
4: Using data $\{\left(S_t^i, A_t^i, R_t^i, S_{t+1}^i\right) : k(i,t) = -k'\}$:
  Compute pseudo-outcomes $\{\tilde{Y}_t^{(i)}(\hat{Z}_t^{1-\tau,k'}, \hat{\overline{Q}}_{t+1}^{-k'})\}_{i=1}^n$ as in eq. (14).
  Fit $\hat{\overline{Q}}_t^{-k'}$ via least-squares regression of $\tilde{Y}_t^{(i)}$ against $(s_t^{(i)}, a_t^{(i)})$.
5: Obtain the robust Q-function by averaging across folds: $\hat{\overline{Q}}_t = \sum_{k'=1}^K \hat{\overline{Q}}_t^{(k)}$
6: Compute $\pi_t^*(s) \in \arg\max_a \hat{\overline{Q}}_t(s,a)$.
7: **end for**

In the main text, we described sample splitting but omitted it from the algorithmic description for a simpler presentation. In Algorithm 3 we discuss the cross-fitting in detail. We use cross-time fitting and introduce folds that partition trajectories and timesteps $k(i,t)$. For $K=2$ we consider timesteps interleaved by parity (e.g. odd/even timesteps in the same fold). We let $-k(i,t)$ denote that nuisance $\hat{\mu}^{-k(i,t)}$ is learned from $\{S_{t'}^{(i)}, A_{t'}^{(i)}, S_{t'+1}^{(i)}\}_{i\in\mathcal{I}_{k(i)}}$, where t' and t have the same parity, e.g. from the $-k(i)$ trajectories and from timesteps of the same evenness or oddness but is only used for evaluation in the other fold.

B.2 Infinite-horizon results

Results for the infinite-horizon setting follow readily from our analysis of the finite-horizon setting and characterization of the uncertainty set. For completeness we state results here, succinctly. First, the algorithm is analogous except with K iterations (restated in Algorithm 4).

In the infinite-horizon setting, we assume the data is generated from the distribution $\mu \in \Delta(\mathcal{S}\times\mathcal{A})$. We instead assume concentratability with respect to stationary distributions.

Assumption 13 (Infinite-Horizon concentratability coefficient). *We assume that there exists $C < \infty$ s.t. for any admissible ν,*

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \frac{\nu(s,a)}{\mu(s,a)} \leq C$$

We first list some helpful lemmas (i.e. infinite-horizon counterparts of the finite-horizon versions).

Our analysis as in Theorem 1 can also be applied to the infinite-horizon case via alternative lemmas standard in the infinite-horizon setting; below we use results from [Chen and Jiang, 2019]. We introduce a discount factor, $\gamma < 1$.

Algorithm 4 Confounding-Robust Fitted-Q-Iteration (Infinite Horizon)

Estimate the marginal behavior policy $\pi^b(a|s)$.
Compute $\{\alpha_k(s^{(i)}, a^{(i)})\}_{i=1}^n$ as in Equation (7).
Initialize $\hat{\overline{Q}}_k = 0$.
for $k = 1, \ldots, K$ **do**
  Compute the nominal outcomes $\{Y_k^{(i)}(\hat{\overline{Q}}_{k-1})\}_{i=1}^n$ as in Equation (11).
  Fit $\hat{Z}_k^{1-\tau}(s, a)$ the $(1-\tau)$th conditional quantile of the outcomes $Y_k^{(i)}$.
  Compute pseudooutcomes $\{\tilde{Y}_k^{(i)}(\hat{Z}_k^{1-\tau}, \hat{\overline{Q}}_{k-1})\}_{i=1}^n$ as in Equation (14).
  Fit $\hat{\overline{Q}}_k$ via least-squares regression of $\tilde{Y}_k^{(i)}$ against $(s^{(i)}, a^{(i)})$.
  Compute $\pi_k^*(s) \in \arg\max_a \hat{\overline{Q}}_k(s, a)$.
end for

Theorem 3 (Infinite-horizon FQI convergence)**.** *Suppose Assumptions 3, 4, 7 and 13 and let* $\bar{V}_{\max} = \frac{1}{1-\gamma} B_R$ *be the upper bound on* $\bar{V}$*. Then, with probability* $> 1 - \delta$*, under Assumption 9, we have that*

$$\left\| \hat{\overline{Q}}_k - \overline{Q}^\star \right\|_{2,\nu} \leq \frac{1-\gamma^k}{1-\gamma} \sqrt{C\left(\epsilon_1 + \epsilon_{\mathcal{Q},\mathcal{Z}}\right)} + \gamma^k \bar{V}_{\max} + o_p(\gamma^k n^{-\frac{1}{2}}).$$

where

$$\epsilon_1 = \frac{56 \bar{V}_{\max}^2 \log\{N(\epsilon, \mathcal{Q}, \|\cdot\|) N(\epsilon, \mathcal{Z}, \|\cdot\|)/\delta\}}{3n} + \sqrt{\frac{32 \bar{V}_{\max}^2 \log\{N(\epsilon, \mathcal{Q}, \|\cdot\|) N(\epsilon, \mathcal{Z}, \|\cdot\|)/\delta\}}{n} \epsilon_{\mathcal{Q},\mathcal{Z}}}.$$

.

C Additional discussion

C.1 Related Work

Connections to pessimism in offline RL. Pessimism is an important algorithmic design principle for offline RL in the *absence* of unobserved confounders [Xie et al., 2021a, Rashidinejad et al., 2021, Jin et al., 2021]. Therefore, robust FQI with lower-confidence-bound-sized Λ gracefully degrades to a pessimistic offline RL method if unobserved confounders were, contrary to our method's use case, not actually present in the data. Conversely, pessimistic offline RL with *state-wise* lower confidence bounds confers some robustness against unobserved confounders. But state-wise LCBs are viewed as overly conservative relative to a profiled lower bound on the average value [Xie et al., 2021a].

C.2 Derivation of the Closed-Form for the Robust Bellman Operator

Proof. Proof of Proposition 5 Dorn et al. [2021] show that the linear program in Proposition 4 has a closed-form solution corresponding to adversarial weights:

$$\tilde{Y}_{f,t}^{-}(s,a) = \mathbb{E}_{\pi^b}\left[W_t^* Y_t | S_t = s, A_t = a\right] \text{ where } W_t^* = \alpha_t \mathbb{I}\left[Y_t > Z_t^{1-\tau}\right] + \beta_t \mathbb{I}\left[Y_t \leq Z_t^{1-\tau}\right].$$

We can derive the form in Proposition 5 with a few additional transformations. Define:

$$\mu_t(s,a) \coloneqq \mathbb{E}_{\pi^b}[Y_t|S_t = s, A_t = a], \quad \mathrm{CVaR}_t^{1-\tau}(s,a) \coloneqq \frac{1}{1-\tau}\mathbb{E}_{\pi^b}\left[Y_t \mathbb{I}\left[Y_t < Z_t^{1-\tau}\right]|S_t = s, A_t = a\right].$$

We use the following identity for any random variables Y and X:

$$\mathbb{E}[Y|X] = \mathbb{E}[Y\mathbb{I}\left[Y > Z^{1-\tau}(Y|X)\right]|X] + \mathbb{E}[Y\mathbb{I}\left[Y \leq Z^{1-\tau}(Y|X)\right]|X]$$

to deduce that

$$\tilde{Y}^{-}_{f,t}(s,a) = \alpha_t\mu_t(s,a) + (\beta_t - \alpha_t)(1-\tau)\mathrm{CVaR}_t^{1-\tau}(s,a),$$

which gives the desired convex combination by noticing that $(\beta_t - \alpha_t)(1-\tau) = (1-\alpha_t)$. □

C.3 Doubly robust estimation of the policy value

We briefly describe an estimation strategy for the off-policy value when we seek to estimate the average policy value (rather than recover the entire $\overline{Q}-$function as we do in this paper). We do so in order to highlight the qualitative differences in estimation. See Jiang and Li [2016], Thomas et al. [2015] for references for off-policy evaluation.

Recall the backward-recursive derivation of off-policy evaluation as in standard presentations of off-policy evaluation for (non-Markovian) MDPs, as in Jiang and Li [2016]. Let $\hat{V}_t^{-}$ denote the robust marginal off-policy value (we superscript by $-$ to demarcate this from the s-conditional value function we study in the rest of the paper). To summarize the derivation intuitively, consider single-timestep robust IPW identification of $\hat{V}_T^{-} = \sum_{a\in\mathcal{A}} \mathbb{E}\left[\frac{\pi^e(A_T|S_T)}{\pi^{b-}(A_T|S_T)}R_T\right] = \inf_{\pi^b\in\mathcal{U}} \mathbb{E}\left[\frac{\pi^e(A_T|S_T)}{\pi^b(A_T|S_T)}R_T\right]$ by inverse-propensity weighting by the robust counterpart $\pi_t^{b-}(a_t \mid s_t)$ of the inverse propensity score. The backwards-recursion proceeds by identifying $\hat{V}_t^{-} = \inf_{\pi^b\in\mathcal{U}}\left\{\sum_{a\in\mathcal{A}} \mathbb{E}[\frac{\pi^e(A_t|S_t)}{\pi^b(A_t|S_t)}(R_t + \hat{V}_{t+1}^{-})]\right\}$, where notably $\hat{V}_{t+1}^{-}$ is a *scalar*. (The backward-recursive derivation is discussed also in Namkoong et al. [2020], Bruns-Smith [2021]).

We first remark that sequential exogenous confounders result in a *time-rectangular* robust decision problem, so that robust backward induction yields a decision rule without issues of time-homogenous or time-inhomogenous preferences (see Delage and Iancu [2015] for a discussion of these challenges otherwise).

For the purpose of highlighting differences in estimation we briefly discuss estimation of the off-policy value based on adversarial propensity weighting. Let $Z_{1-\tau}^{u}$ denote the quantile function of some u (the time indexing will be clear from context). Unbiasedness for time $t = 0$ follows directly by backwards induction.

Proposition 10.

$$V_t^{-} = \sum_a \mathbb{E}\left[\pi(a \mid s)(R_t + \hat{V}_{t+1}^{-})(a_t\mathbb{I}\left[R_t > Z_{1-\tau}^{R_t}\right] + b_t\mathbb{I}\left[R_t < Z_{1-\tau}^{R_t}\right])\right] \tag{19}$$

where

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

Let $(\cdot)^*$ denote a well-specified nuisance function and $(\cdot)^\dagger$ denote a mis-specified nuisance function. Let $\overline{Q}$ be attained via robust FQE. Define an estimate with an additional control variate:

$$\psi(\pi, Z, \overline{Q}_t) = \frac{\mathbb{I}\left[\pi^e = a\right]}{\pi^{b-}}(R_t^+ + \hat{V}_{t+1}^{DR}) + \left\{1 - \frac{\mathbb{I}\left[\pi^e = a\right]}{\pi^{b-}}\right\}\overline{Q}_t(s,a). \tag{20}$$

Then

$$\mathbb{E}[\psi(\pi^*, Z^*, \overline{Q}_t^\dagger)] = \mathbb{E}[\psi(\pi^*, Z^\dagger, \overline{Q}_t^*)] = V_t^-.$$

Equation (19) is notable because it shows how the assumption of sequentially exogenous unobserved confounders (Assumption 1) leads to off-policy evaluation that (sequentially) evaluates single-step quantile functions. Moreover, this is qualitatively different than what arises in the fitted-Q-evaluation setting. The proof of eq. (19) follows directly from previous results on the closed-form solution (we follow a representation of [Tan, 2022] for convenience) and linear equivariance to a constant shift of quantile regression. More specifically, because

$$Z_\tau^{R_t + v - q_t(s,a)}(s,a) = Z_\tau^{R_t}(s,a) + v - q_t(s,a) \tag{21}$$

for any scalar constant v and deterministic function $q(s,a)$. (Also note that a_t, b_t differ from α_t, β_t used in the main text by a multiplicative factor of $\pi_t^b(a|s)$ because we evaluate marginal policy values, rather than optimize with respect to the (s,a)-conditional distribution).

Proof. Proof of Proposition 10

Start from robust inverse propensity weighting and add the control variate $\mathbb{E}\left[\left\{1 - \frac{\mathbb{I}[\pi^e = a]}{\pi^{b-}}\right\}\overline{Q}_t(s,a)\right]$, where $\overline{Q}_t$ is obtained via robust FQE:

$$\epsilon_t^- = R_t + \hat{V}_{t+1}^- - \overline{Q}_t, \qquad \hat{V}_t^- = \mathbb{E}\left[\overline{Q}_t + \frac{\mathbb{I}\left[A_t = a\right]}{\pi^{b-}}\epsilon_t^-\right]$$

Expanding out the adversarial propensity and applying quantile equivariance from Equation (21):

$$\begin{aligned}
\hat{V}_t^- &= \mathbb{E}\left[\overline{Q}_t + A\epsilon_t^-/\pi - (\Lambda - \Lambda^{-1})A\frac{1-\pi}{\pi}\left\{\left((1-\tau)(\epsilon_t^- - Z_{1-\tau}^{\epsilon_t^-}(X,1))_+ + \tau(\epsilon_t^- - Z_{1-\tau}^{\epsilon_t^-}(X,1))_-\right)\right\}\right] \\
&= \mathbb{E}\left[\overline{Q}_t + A\epsilon_t^-/\pi - (\Lambda - \Lambda^{-1})A\frac{1-\pi}{\pi}\left\{\left((1-\tau)(R_t - Z_{1-\tau}^{R_t}(X,1))_+ + \tau(R_t - Z_{1-\tau}^{R_t}(X,1))_-\right)\right\}\right]
\end{aligned}$$

Next we verify the proposed multiple robustness properties. Due to the functional form of π_*^{b-}, we use $*, \dagger$ specification notation on π^{b-} to refer jointly to (π^*, Z^*) or $(\pi^\dagger, Z^\dagger)$. When we have $(\pi^*, Z^*, \overline{Q}_t^\dagger)$:

$$\begin{aligned}
\mathbb{E}[\psi^{DR}(\pi^*, Z^*, \overline{Q}_t^\dagger, \eta_{\overline{Q}_t}^\dagger)] &= \mathbb{E}\left[\frac{\mathbb{I}\left[\pi^e = a\right]}{\pi_*^{b-}}(R_t^+ + \hat{V}_{t+1}^{DR}) + \left\{1 - \frac{\mathbb{I}\left[\pi^e = a\right]}{\pi_*^{b-}}\right\}\overline{Q}_t^\dagger\right] && (22) \\
&= V_t^-. && (23)
\end{aligned}$$

Under assumption of well-specified π^*, Z^*, we have that $\mathbb{E}\left[\frac{\mathbb{I}[\pi^e=a]}{\pi_*^{b-}}(R_t^+ + \hat{V}_{t+1}^{DR})\right] = V_t$. By iterated expectation, $\mathbb{E}\left[\left\{1 - \frac{\mathbb{I}[\pi^e=a]}{\pi_*^{b-}}\right\}\overline{Q}_t^\dagger\right] = 0$. Also,

$$
\begin{aligned}
&\mathbb{E}[\psi^{DR}(\pi^*, Z^\dagger, \overline{Q}_t^*, \eta_{\overline{Q}_t}^\dagger)] \\
&= \mathbb{E}\left[\frac{\mathbb{I}\left[\pi^e = a\right]}{\pi_\dagger^{b-}}(R_t^+ + \hat{V}_{t+1}^{DR}) + \left\{1 - \frac{\mathbb{I}\left[\pi^e = a\right]}{\pi_*^{b-}}\right\}\overline{Q}_t^*\right] \\
&= \mathbb{E}\left[\overline{Q}_t^* + \frac{\mathbb{I}\left[\pi^e = a\right]}{\pi_\dagger^{b-}}(R_t^+ + \hat{V}_{t+1}^{DR} - \overline{Q}_t^*)\right] \\
&= V_t^-.
\end{aligned}
$$

The second term is 0 by iterated expectation. $\square$

D Proofs for Robust FQE/FQI

D.0.1 Auxiliary lemmas for robust FQE/FQI

Lemma 7 (Higher-order quantile error terms)**.** *Assume Assumption 4 (i.e. bounded conditional density by M_P), and that $Z_t^{1-\tau}$ is differentiable with respect to s and its gradient is Lipschitz continuous. Then, for $f_t = R_t + \hat{\overline{Q}}_{t+1}$, if $\hat{Z}_t^{1-\tau}$ is $O_p(w_n)$ sup-norm consistent, i.e.* $\sup_{s\in\mathcal{S}} |Z_t^{1-\tau} - \hat{Z}_t^{1-\tau}| = O_p(w_n)$*, uniformly over $s \in \mathcal{S}$,*

$$|\mathbb{E}[(f_t - Z_t^{1-\tau})(\mathbb{I}[f_t \leq \hat{Z}_t^{1-\tau}] - \mathbb{I}[f_t \leq Z_t^{1-\tau}]) \mid S = s, A = 1]| = O_p(w_n^2), \tag{24}$$

and

$$\mathbb{E}[(Z_t^{1-\tau} - \hat{Z}_t^{1-\tau}) \left(\mathbb{I}[f \leq Z_t^{1-\tau}] - (1-\tau)\right) \mid A = 1] \leq M_P \mathbb{E}[(Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})^2 \mid A = a]. \tag{25}$$

Lemma 7 is a technical lemma which summarizes the properties of the orthogonalized target which lead to quadratic bias in the first-stage estimation error of $\hat{Z}_t$. Equation (24) is a slight modification of [Olma, 2021]/[Kato, 2012, A.3]; eq. (25) is a slight modification of Semenova [2023, Lemma 4.1].

Lemma 8 (Bernstein concentration for least-squares loss (under approximate realizability))**.** *Suppose Assumption 8 and that:*

1. *Approximate realizability: $\mathcal{Q}$ approximately realizes $\overline{\mathcal{T}}\mathcal{Q}$ in the sense that $\forall f \in \mathcal{Q}, z \in \mathcal{Z}$, let $q_f^\star = \arg\min_{q\in\mathcal{Q}} \|q - \overline{\mathcal{T}} f\|_{2,\mu}$, then $\left\|q_f^\star - \overline{\mathcal{T}} f\right\|_{2,\mu}^2 \leq \epsilon_{\mathcal{Q},\mathcal{Z}}$.*

The dataset $\mathcal{D}$ is generated from P_{obs} as follows: $(s,a) \sim \mu, r = R(s,a), s' \sim P(s' \mid s,a)$. We have that $\forall f \in \mathcal{Q}$, with probability at least $1-\delta$,

$$\mathbb{E}_\mu[\ell(\widehat{\mathcal{T}}_{\mathcal{Z}} f; f)] - \mathbb{E}_\mu[\ell(g_f^\star; f)] \leq \frac{56 V_{\max}^2 \ln \frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{3n} + \sqrt{\frac{32 V_{\max}^2 \ln \frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{n} \epsilon_{\mathcal{Q},\mathcal{Z}}}$$

Lemma 9 (Stability of covering numbers)**.** *We relate the covering numbers of the squared loss function class, denoted as $\mathcal{L}_{q(z'),z}(q_{t+1})$, to the covering numbers of the function classes $\mathcal{Q}, \mathcal{Z}$. Define the squared loss function class as:*

$$\mathcal{L}_{q(z'),z}(q_{t+1}) = \left\{\ell(q(z'), q_{t+1}; z) - \ell(\overline{Q}_{t,Z_t}^\dagger, q_{t+1}; z) \colon q(z') \in \{\mathcal{Q} \otimes \mathcal{Z}\}, z \in \mathcal{Z}\right\}$$

Then

$$N_{[]}\left(2\epsilon L, \mathcal{L}_{q(z'),z}, \|\cdot\|\right) \leq N(\epsilon, \mathcal{Q} \times \mathcal{Z}, \|\cdot\|).$$

Lemma 10 (Difference of indicator functions)**.** *Let $\widehat{f}$ and f take any real values. Then* $\left|\mathbb{I}[\widehat{f} > 0] - \mathbb{I}\left[f > 0\right]\right| \leq \mathbb{I}[|f| \leq |\widehat{f} - f|]$

D.1 Proofs of theorems

Proof. Proof of Theorem 1

The squared loss (with respect to a given conditional quantile function Z) is:

$$\ell(q, q_{t+1}; Z)$$
$$= \left(\alpha(R + q_{t+1}) + (1-\alpha)\Big(Z_t^{1-\tau} + \frac{1}{1-\tau}\left((R + q_{t+1} - Z_t^{1-\tau})_- - Z_t^{1-\tau} \cdot (\mathbb{I}\left[R + q_{t+1} \leq Z_t^{1-\tau}\right] - (1-\tau)) \right) \Big) - q_t \right)^2$$

We let $\hat{Z}_{t,Q_{t+1}}$ and $Z_{t,Q_{t+1}}$ denote estimated and oracle conditional quantile functions, respectively, with respect to a target function that uses Q_{t+1} estimate. Where the next-timestep Q function is fixed (as it is in the following analysis) we drop the Q_{t+1} from the subscript.

Define

$$\hat{\overline{Q}}_{t,Z_t} \in \arg\min_q \mathbb{E}_n[\ell(q, \hat{\overline{Q}}_{t+1}; Z_t)]$$

and for $z \in \{\hat{Z}_t, Z_t\}$, define the following *oracle* Bellman error projections $\overline{Q}^{\dagger}_{t,z}$ of the iterates of the algorithm:

$$\overline{Q}^{\dagger}_{t,z} = \arg\min_{q_t \in \mathcal{Q}_t} \| q_t - \overline{\mathcal{T}}^*_{t,z} \hat{\overline{Q}}_{t+1} \|_{\mu_t}.$$

Relating the Bellman error to FQE loss. The bias-variance decomposition implies if U, V are conditionally uncorrelated given W, then

$$\mathbb{E}[(U - V)^2 \mid W] = \mathbb{E}[(U - \mathbb{E}[V \mid W])^2 \mid W] + Var[V \mid W].$$

Hence a similar relationship holds for the robust Bellman error as for the Bellman error:

$$\mathbb{E}[\ell(q, \overline{Q}_{t+1}; Z)^2] = \| q - \overline{\mathcal{T}}^* \overline{Q}_{t+1} \|_{\mu} + Var[W_t^{*,\pi}(Z)(R_t + \bar{V}_{\overline{Q}_{t+1}}(S_{t+1})) \mid S_t, A].$$

which is used to decompose the Bellman error as follows:

$$\|\hat{\overline{Q}}_{t,\hat{Z}_t} - \overline{\mathcal{T}}^*_{t,Z_t}\hat{\overline{Q}}_{t+1}\|^2_{\mu_t} = \mathbb{E}_\mu[\ell(\hat{\overline{Q}}_{t,\hat{Z}_t}, \hat{\overline{Q}}_{t+1}; Z_t)] - \mathbb{E}_\mu[\ell(\overline{Q}^{\dagger}_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] + \|\overline{Q}^{\dagger}_{t,Z_t} - \overline{\mathcal{T}}^*_t \hat{\overline{Q}}_{t+1}\|^2_{\mu_t}.$$

Then,

$$\begin{aligned}
&\|\hat{\overline{Q}}_{t,\hat{Z}_t} - \overline{\mathcal{T}}^*_{t,Z_t}\hat{\overline{Q}}_{t+1}\|^2_{\mu_t} \\
&= \mathbb{E}_\mu[\ell(\hat{\overline{Q}}_{t,\hat{Z}_t}, \hat{\overline{Q}}_{t+1}; Z_t)] - \mathbb{E}_\mu[\ell(\hat{\overline{Q}}_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] && (26)\\
&+ \mathbb{E}_\mu[\ell(\hat{\overline{Q}}_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] - \mathbb{E}_\mu[\ell(\overline{Q}^{\dagger}_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] && (27)\\
&+ \|\overline{Q}^{\dagger}_{t,Z_t} - \overline{\mathcal{T}}^*_t \hat{\overline{Q}}_{t+1}\|^2_{\mu_t} && (28)
\end{aligned}$$

We bound eq. (26) by orthogonality and eq. (27) by Bernstein inequality arguments.

We bound the first term. Let f denote the Bellman residual. Let $x = f$, $(a - x) = Q - f$, $b = Q'$. Since, by expanding the square and Cauchy-Schwarz, we obtain the following elementary inequality:

$$\begin{aligned}
(a-x)^2 - (b-x)^2 &= (a-b)^2 + 2(a-b)(b-x) \\
&\leq (a-b)^2 + \sqrt{\mathbb{E}[(a-b)^2]\mathbb{E}[(b-x)^2]}
\end{aligned}$$

Applying the above, we have that

$$\mathbb{E}_\mu[\ell(\hat{\overline{Q}}_{t,Z_t},\hat{\overline{Q}}_{t+1};Z_t)] - \mathbb{E}_\mu[\ell(\overline{Q}^\dagger_{t,Z_t},\hat{\overline{Q}}_{t+1};Z_t)] \leq \underbrace{\|(\hat{\overline{Q}}_{t,Z_t} - \overline{Q}^\dagger_{t,Z_t})\|_2^2}_{o_p(n^{-1}) \text{ by } \textit{Proposition } 6} + \|(\hat{\overline{Q}}_{t,Z_t} - \overline{Q}^\dagger_{t,Z_t})^2\| \underbrace{\|\hat{\overline{Q}}_{t,Z_t} - Y_t(\hat{\overline{Q}}_{t+1};Z_t)\|}_{=O_p(n^{-1/2}) \text{ by realizability}}$$

Therefore

$$\mathbb{E}_\mu[\ell(\hat{\overline{Q}}_{t,Z_t},\hat{\overline{Q}}_{t+1};Z_t)] - \mathbb{E}_\mu[\ell(\overline{Q}^\dagger_{t,Z_t},\hat{\overline{Q}}_{t+1};Z_t)] = o_p(n^{-1}).$$

We bound eq. (27) by Lemma 8 directly.

Supposing Assumption 8, we obtain that

$$\left\|\hat{Q}_t - \overline{\mathcal{T}}^\star_t \hat{Q}_{t+1}\right\|^2_{\mu_t} \leq \epsilon_{\mathcal{Q},\mathcal{Z}} + \frac{56V^2_{\max}\ln\frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{3n} + \sqrt{\frac{32V^2_{\max}\ln\frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{n}\epsilon_{\mathcal{Q},\mathcal{Z}}} + o_p(n^{-1}).$$

Instead, supposing Assumption 9, instantiate the covering numbers choosing $\epsilon = O(n^{-1})$. Lemma 9 bounds the bracketing numbers of the (Lipschitz over a bounded domain) loss function class with the covering numbers of the primitive function classes $\mathcal{Q},\mathcal{Z}$. Supposing that Bellman completeness holds with respect to $\mathcal{Q},\mathcal{Z}$, approximate Bellman completeness holds over the ϵ-net implied by the covering numbers with $\epsilon_{\mathcal{Q},\mathcal{Z}} = O(n^{-1})$ and we obtain that:

$$\begin{aligned}\left\|\hat{Q}_t - \overline{\mathcal{T}}^\star_t \hat{Q}_{t+1}\right\|^2_{\mu_t} &\leq \epsilon_{\mathcal{Q},\mathcal{Z}} + \frac{56V^2_{t,\max}\log\{N(\epsilon,\mathcal{Q},\|\cdot\|)N(\epsilon,\mathcal{Z},\|\cdot\|)/\delta\}}{3n} \\ &\quad + \sqrt{\frac{32V^2_{t,\max}\log\{N(\epsilon,\mathcal{Q},\|\cdot\|)N(\epsilon,\mathcal{Z},\|\cdot\|)/\delta\}}{n}\epsilon_{\mathcal{Q},\mathcal{Z}}} + o_p(n^{-1}). \\ &\leq \epsilon_{\mathcal{Q},\mathcal{Z}} + \frac{56V^2_{t,\max}\log\{N(\epsilon,\mathcal{Q},\|\cdot\|)N(\epsilon,\mathcal{Z},\|\cdot\|)/\delta\}}{3n}\end{aligned}$$

□

Proof. Proof of Theorem 3 Note that Lemma 13, [Chen and Jiang, 2019] establishes the Bellman error as an upper bound to the policy suboptimality. It states: Let $f : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ and $\hat{\pi} = \pi_f$ be the policy of interest, we have

$$\bar{V}^\star - \bar{V}^{\hat{\pi}} \leq \sum_{t=1}^\infty \gamma^{t-1}\left(\left\|\bar{Q}^\star - f\right\|_{2,\eta_t^{\hat{\pi}}\times\pi^\star} + \left\|\bar{Q}^\star - f\right\|_{2,\eta_t^{\hat{\pi}}\times\hat{\pi}}\right).$$

Choosing $f = \hat{\bar{Q}}_k$ and $f' = \hat{\bar{Q}}_{k-1}$ in [Chen and Jiang, 2019, Lemma 15] gives

$$\left\|\hat{\bar{Q}}_k - \bar{Q}^\star\right\|_{2,\nu} \leq \sqrt{C}\left\|\hat{\bar{Q}}_k - \mathcal{T}\hat{\bar{Q}}_{k-1}\right\|_{2,\mu} + \gamma\left\|\hat{\bar{Q}}_{k-1} - \bar{Q}^\star\right\|_{2,P(\nu)\times\pi_{\hat{\bar{Q}}_{k-1},\bar{Q}^\star}}. \tag{29}$$

Note that we can apply the same analysis with $P(\nu)\times\pi_{\hat{\bar{Q}}_{k-1},\bar{Q}^\star}$ replacing the ν distribution on the left hand side, and expand the inequality k times. Then it remains to upper bound

$\left\|\hat{\overline{Q}}_k - \mathcal{T}\hat{\overline{Q}}_{k-1}\right\|_{2,\mu}$, which we can do via the same analysis of eqs. (26) to (28). Following the analysis of the proof of Theorem 1, we then obtain, with probability $\geq 1-\delta$,

$$\left\|\hat{\overline{Q}}_k - \overline{\mathcal{T}}_t^{\star}\hat{\overline{Q}}_{k-1}\right\|_{\mu_t}^2 \leq \epsilon_{\mathcal{Q},\mathcal{Z}} + \epsilon_1 + o_p(n^{-1}),$$

where

$$\epsilon_1 = \frac{56 V_{t,\max}^2 \log\{N(\epsilon, \mathcal{Q}, \|\cdot\|)N(\epsilon, \mathcal{Z}, \|\cdot\|)/\delta\}}{3n} + \sqrt{\frac{32 V_{\max}^2 \log\{N(\epsilon, \mathcal{Q}, \|\cdot\|)N(\epsilon, \mathcal{Z}, \|\cdot\|)/\delta\}}{n}\epsilon_{\mathcal{Q},\mathcal{Z}}}.$$

Since ϵ_1 and $\epsilon_{\mathcal{Q},\mathcal{Z}}$ are independent of k, and the bound holds uniformly over k, we have that, plugging the above back into the recursive expansion of Equation (29):

$$\left\|\hat{\overline{Q}}_k - \overline{Q}^{\star}\right\|_{2,\nu} \leq \frac{1-\gamma^k}{1-\gamma}\sqrt{C\left(\epsilon_1 + \epsilon_{\mathcal{Q},\mathcal{Z}}\right)} + \gamma^k \bar{V}_{\max}.$$

□

D.2 Proofs of intermediate results

D.2.1 Orthogonality

Proof. Proof of Proposition 6 We first focus on the case of a single action, $a=1$. First recall that in the population, $\mathbb{E}[Z_t^{1-\tau} + \frac{1}{1-\tau}(f_t - Z_t^{1-\tau}) \mid s, a] = \frac{1}{1-\tau}\mathbb{E}[f_t \mathbb{I}\left[f_t \leq Z_t^{1-\tau}\right] \mid s, a]$. In the analysis below we study this truncated conditional expectation representation.

$$\|\hat{\overline{Q}}_t(S,1) - \overline{Q}_t(S,1)\| \lesssim \|\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, A=1]\| + \|\hat{\overline{Q}}_t(S,1) - \overline{Q}_t(S,1)\|$$
by Prop. 1 of Kennedy [2020] (regression stability)

Prop. 1 of Kennedy [2020] provides bounds on how regression upon pseudooutcomes with estimated nuisance functions relates to the case with known nuisance functions.

It remains to relate $\|\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, A=1]\|$ to the terms comprising the pointwise bias, which are bounded by Lemma 7. We define these terms as:

$$\begin{aligned}
B_1^1(S) &= \mathbb{E}\left[\frac{1-\tilde{\alpha}}{1-\tau}\left\{(f_t - Z_t^{1-\tau})\left(\mathbb{I}\left[f_t \leq \hat{Z}_t^{1-\tau})\right] - \mathbb{I}\left[f_t \leq Z_t^{1-\tau})\right]\right)\right\} \mid S, A=1\right] \\
B_2^1(S) &= \mathbb{E}\left[\frac{1-\tilde{\alpha}}{1-\tau}\left\{(Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})\left(\mathbb{I}\left[f \leq Z_t^{1-\tau}\right] - (1-\tau)\right)\right\} \mid S, A=1\right].
\end{aligned}$$

Lemma 7 bounds these terms as quadratic in the first-stage estimation error of $\hat{Z}_t$.

We have that

$$\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, 1] = B_1^1(S) + B_2^1(S).$$

To see this, note:

$$
\begin{aligned}
&\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, 1] \\
&= \mathbb{E}\left[\frac{1-\tilde{\alpha}}{1-\tau}\left\{\left(f_t \mathbb{I}\left[f_t \leq \hat{Z}_t^{1-\tau})\right] - f_t \mathbb{I}\left[f_t \leq Z_t^{1-\tau})\right]\right)\right.\right. \\
&\quad \left.\left. - \left(\hat{Z}_t^{1-\tau} \cdot (\mathbb{I}\left[f \leq \hat{Z}_t^{1-\tau}\right] - (1-\tau)) - Z_t^{1-\tau} \cdot (\mathbb{I}\left[f \leq Z_t^{1-\tau}\right] - (1-\tau))\right) \pm Z_t^{1-\tau} \cdot \mathbb{I}\left[f \leq \hat{Z}_t^{1-\tau}\right]\right\} \mid S, A=1\right] \\
&= \mathbb{E}\left[\frac{1-\tilde{\alpha}}{1-\tau}\left\{(f_t - Z_t^{1-\tau})\mathbb{I}\left[f_t \leq \hat{Z}_t^{1-\tau})\right] - (f_t - Z_t^{1-\tau})\mathbb{I}\left[f_t \leq Z_t^{1-\tau})\right]\right.\right. \\
&\quad \left.\left. + (Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})\mathbb{I}\left[f \leq Z_t^{1-\tau}\right] - (Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})(1-\tau)\right\} \mid S, A=1\right] \\
&= \mathbb{E}\left[\frac{1-\tilde{\alpha}}{1-\tau}\left\{(f_t - Z_t^{1-\tau})\left(\mathbb{I}\left[f_t \leq \hat{Z}_t^{1-\tau})\right] - \mathbb{I}\left[f_t \leq Z_t^{1-\tau})\right]\right) + (Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})\left(\mathbb{I}\left[f \leq Z_t^{1-\tau}\right] - (1-\tau)\right)\right\} \mid S, A = 1\right] \\
&= B_1^1(S) + B_2^1(S)
\end{aligned}
$$

Finally, we relate the root mean-squared conditional bias,

$$
\|\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, A=1]\|,
$$

to the above quadratic error as follows. Using the inequalities $(a+b)^2 \leq 2(a^2+b^2)$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ (for nonnegative a, b), we obtain that

$$
\begin{aligned}
\|\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, A=1]\| &= \sqrt{\mathbb{E}[(B_1^1(S) + B_2^1(S))^2 \mid A=1]} \\
&\leq \sqrt{\mathbb{E}[2\{(B_1^1(S))^2 + (B_2^1(S))^2\} \mid A=1]} \\
&\leq \sqrt{2\mathbb{E}[(B_1^1(S))^2 \mid A=1]} + \sqrt{2\mathbb{E}[(B_2^1(S))^2 \mid A=1]}.
\end{aligned}
$$

The result follows by the uniform bounds of Lemma 7.

□

Proof. Proof of Lemma 7 Proof of eq. (24):

For $l > 0$, define

$$
\mathcal{M}_n^a(l) = \left\{g : \mathcal{S} \to \mathbb{R} \text{ s.t. } \sup_{s \in \mathcal{S}} |g(s) - Z_t^{1-\tau}(s,a)| \leq l w_n\right\}
$$

Define

$$
U_n(g,s) := |\mathbb{E}[(f_t - Z_t^{1-\tau})(\mathbb{I}\left[f_t \leq \hat{Z}_t^{1-\tau}\right] - \mathbb{I}\left[f_t \leq Z_t^{1-\tau}\right]) \mid S=s, A=1]|
$$

We will show that for every $l > 0, s \in \mathcal{S}$:

$$
\sup_{g \in \mathcal{M}_n(l)} U_n(g,s) = O_p\left(w_n^2\right)
$$

Breaking up the absolute value,

$$
U_n(g,s) \leq \mathbb{E}[(f_t - Z_t^{1-\tau})(\mathbb{I}\left[Z_t^{1-\tau} \leq f_t \leq g\right]) \mid S=s, A=1] + \mathbb{E}[(Z_t^{1-\tau} - f_t)(\mathbb{I}\left[g \leq f_t \leq Z_t^{1-\tau}\right]) \mid S=s, A=1]
$$

We will bound the first term, bounding the second term is analogous. Define

$$U_{1,n}(g,s) \coloneqq \mathbb{E}[(f_t - Z_t^{1-\tau})(\mathbb{I}\left[Z_t^{1-\tau} \leq f_t \leq g\right]) \mid S = s, A = 1]$$

Observe that

$$\begin{aligned}\sup_{g \in \mathcal{M}_n(l)} U_{1,n}(g,s) &= \mathbb{E}[(f_t - Z_t^{1-\tau})(\mathbb{I}\left[Z_t^{1-\tau} \leq f_t \leq Z_t^{1-\tau} + l w_n\right]) \mid S = s, A = 1] \\ &\leq M_P l^2 w_n^2\end{aligned}$$

The result follows.

Proof of eq. (25):

The argument follows that of Semenova [2017]. The difference of indicators is nonzero on the events:

$$\begin{aligned}\mathcal{E}^- &\coloneqq \{f_t - \hat{Z}_t^{1-\tau} < 0 < f_t - Z_t^{1-\tau}\} \\ \mathcal{E}^+ &\coloneqq \{f_t - Z_t^{1-\tau} < 0 < f_t - \hat{Z}_t^{1-\tau}\}\end{aligned}$$

On these events, the estimation error upper bounds the exceedance

$$\{\mathcal{E}^- \cup \mathcal{E}^+\} \implies \{|Z - f| < |Z_t^{1-\tau} - \hat{Z}_t^{1-\tau}|\} \tag{30}$$

(since $\mathcal{E}^- \implies \{f - \hat{Z}_t^{1-\tau} < 0 < f - Z_t^{1-\tau}\}$ and $\mathcal{E}^+ \implies \{0 < Z_t^{1-\tau} - f < Z_t^{1-\tau} - \hat{Z}_t^{1-\tau}\}$.)

Then

$$\begin{aligned}\mathbb{E}[(f_t - Z_t^{1-\tau})\mathbb{I}\left[\mathcal{E}^- \cup \mathcal{E}^+\right] \mid S = s, A = 1] &= \int_{-|Z_t^{1-\tau} - \hat{Z}_t^{1-\tau}|}^{|Z_t^{1-\tau} - \hat{Z}_t^{1-\tau}|} (f_t(s,a,s') - Z_t^{1-\tau}) P(s' \mid s,a) ds' \\ &\leq M_P \mathbb{E}[(Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})^2 \mid S = s, A = 1]\end{aligned}$$

Assumption 6 ensures the result holds for state distributions that could arise during policy fitting. The above results hold conditionally on some action $A = 1$ but hold for all actions.

□

D.2.2 Other lemmas

Proof. Proof of Lemma 8 Recall that

$$\ell(q, q_{t+1}; Z)$$
$$= \left(\alpha(R + q_{t+1}) + (1-\alpha)\left(Z_t^{1-\tau} + \frac{1}{1-\tau}\left((R + q_{t+1} - Z_t^{1-\tau})_- - Z_t^{1-\tau} \cdot (\mathbb{I}\left[R + q_{t+1} \leq Z_t^{1-\tau}\right] - (1-\tau))\right)\right) - q_t\right)^2$$

Define $f_{q',z} =$ Define X to be the difference of the integrands.

Step 1:

$$Var(X(g,f,z,g_f^*)) \leq 4V_{\max}^2 \|\hat{\overline{Q}}_{t,Z_t} - \overline{Q}_{t,Z_t}^{\dagger}\|_2^2$$

(by similar arguments as in the original paper). By the same arguments (i.e. adding and subtracting $\overline{\mathcal{T}} f$) we obtain that

$$\|\hat{\overline{Q}}_{t,Z_t} - \overline{Q}_{t,Z_t}^{\dagger}\|_2^2 \leq 2(\mathbb{E}[X(g,f,z,g_f^*)] + 2\epsilon_{\mathcal{Q},\mathcal{Z}})$$

Therefore,

$$Var(X(g,f,z,g_f^*)) \leq 8V_{\max}^2(\mathbb{E}[X(g,f,z,g_f^*)] + 2\epsilon_{\mathcal{Q},\mathcal{Z}}).$$

Applying (one-sided) Bernstein's inequality uniformly over $\mathcal{Q}, \mathcal{Z}$, we obtain:

$$\begin{aligned}&\mathbb{E}\left[X(g,f,z,g_f^*)\right] - \mathbb{E}_n[X(g,f,z,g_f^\star)]\\ &\leq \sqrt{\frac{16V_{\max}^2\left(\mathbb{E}\left[X(g,f,z,g_f^\star)\right] + 2\epsilon_{\mathcal{F},\mathcal{Z}}\right)\ln\frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{n}} + \frac{4V_{\max}^2\ln\frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{3n}\end{aligned}$$

Note that $\hat{\overline{Q}}_{t,Z_t}$ minimizes both $\mathbb{E}_n[\ell(q,\hat{\hat{Q}}_{t+1};Z_t)]$ and $\mathbb{E}[(q,\hat{\hat{Q}}_{t+1},Z_t,\overline{Q}^*_{\hat{\hat{Q}}_{t+1}})]$ with respect to q. Therefore, by completeness since the Bayes-optimal predictor is realizable,

$$\mathbb{E}_n[\ell(\hat{\overline{Q}}_{t,Z_t},\hat{\hat{Q}}_{t+1};Z_t)] \leq \mathbb{E}_n[\ell(\overline{Q}^\dagger_{t,Z_t},\hat{\hat{Q}}_{t+1};Z_t)] = 0$$

Therefore (solving for the quadratic formula),

$$\mathbb{E}[X(\hat{\overline{Q}}_{t,Z_t},\hat{Q}_{t+1},Z_t,\overline{Q}^\dagger_{t,Z_t})] \leq \frac{56V_{\max}^2\ln\frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{3n} + \sqrt{\frac{32V_{\max}^2\ln\frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{n}\epsilon_{\mathcal{F},\mathcal{Z}}}$$

□

Proof. Proof of Lemma 9 We show this result by establishing Lipschitz-continuity of the squared loss function class (with respect to the product function class of $\mathcal{Q}\times\mathcal{Z}$).

We use a stability result on the bracketing number under Lipschitz transformation. Classes of functions $x \mapsto f_\theta(x)$ that are Lipschitz in the index parameter $\theta \in \Theta$ have bracketing numbers readily related to the covering numbers of Θ. Suppose that

$$|f_{\theta'}(x) - f_\theta(x)| \leq d(\theta',\theta)F(x),$$

for some metric d on the index set, function F on the sample space, and every x. Then $(\operatorname{diam}\Theta)F$ is an envelope function for the class $\{f_\theta - f_{\theta_0} : \theta\in\Theta\}$ for any fixed θ_0. We invoke Theorem 2.7.11 of van de Vaart and Wellner [1996] which shows that the bracketing numbers of this class are bounded by the covering numbers of Θ.

Theorem 4 ([van de Vaart and Wellner, 1996], Theorem 2.7.11)**.** *Let $\mathcal{F} = \{f_\theta : \theta\in\Theta\}$ be a class of functions satisfying the preceding display for every θ', θ and some fixed function F. Then, for any norm $\|\cdot\|$,*

$$N_{[]}\left(2\epsilon\|F\|,\mathcal{F},\|\cdot\|\right) \leq N(\epsilon,\Theta,d).$$

Let $\mathcal{F} = \{f_\theta : \theta\in\Theta\}$ be a class of functions satisfying the preceding display for every s and θ and some fixed envelope function F. Then, for any norm $\|\cdot\|$,

$$N_{[]}\left(2\epsilon\|F\|,\mathcal{F},\|\cdot\|\right) \leq N(\epsilon,\Theta,d).$$

This shows that the bracketing numbers of the loss function class can be expressed via the covering numbers of the estimated function classes $\mathcal{Q},\mathcal{Z}$, which are the primitive function classes of estimation, for which results are given in various references for typical function classes.

Denote

$$
\begin{aligned}
g(q_{t+1}) &= \alpha(s,a)(R+q_{t+1}) \\
h(z) &= (1-\alpha)\Big(\frac{1}{1-\tau}\left(z+(R+q_{t+1}-z)_- - z\cdot\left(\mathbb{I}\left[R+q_{t+1}\le z\right]-(1-\tau)\right)\right)\Big)
\end{aligned}
$$

and notate

$$\ell(q,q_{t+1};z) = (q - g(q_{t+1}) + h(q_{t+1},z))^2.$$

Note that $\frac{1}{1-\tau} = (1+\Lambda)$. Assuming bounded rewards, define $D_{z,t}, D_{q,t}$ as the diameters of $\mathcal{Q}_t, \mathcal{Z}_t$, respectively and note that $D_{z,t} \approx D_{q,t}$. Note that $h(q_{t+1},z)$ is $(1-\alpha_{\min})(3(1+\Lambda)+1)$-Lipschitz in z (since the sum of Lipschitz continuous functions is Lipschitz) and it is $(1-\alpha_{\min})\left(1+(1+\Lambda)(\frac{D_{z,t}}{D_{q,t}}+1)\right)$-Lipschitz in q_{t+1}. Further, $g(q_{t+1})$ is $\alpha_{\max}$-Lipschitz in q_{t+1}. Therefore, $\ell\left(q,q_{t+1};z\right)$ is $D_{q,t}$ Lipschitz in q, $L^C_{q,t+1}$-Lipschitz in q_{t+1} and $L^C_{z,t}$-Lipschitz in z, with $L^C_{q,t+1}, L^C_{z,t}$ defined as follows:

$$
\begin{aligned}
L^C_{q,t+1} &= (2D_{q,t+1}+D_{z,t})(1-\alpha_{\min})\left(\left(1+(1+\Lambda)(\frac{D_{z,t}}{D_{q,t+1}}+1)\right)+\alpha_{\max}\right) \\
L^C_{z,t} &= (2D_{q,t+1}+D_{z,t})(1-\alpha_{\min})(3(1+\Lambda)+1).
\end{aligned}
$$

Therefore we have shown that restrictions of $\ell\left(q,q_{t+1};z\right)$ to the q_{t+1}, z coordinates are individually Lipschitz. We leverage the fact that a function $f:\mathbb{R}^n\to\mathbb{R}$ is Lipschitz if and only if there exists a constant L such that the restriction of f to every line parallel to a coordinate axis is Lipschitz with constant L. Choosing

$$L_t = \sqrt{3}\max\{D_q, L^C_{q,t+1}, L^C_{z,t}\}$$

gives that $\ell\left(q,q_{t+1};z\right)$ is L_t-Lipschitz.

□

Proof. Proof of Corollary 1 Lemma 9 gives that $\ell\left(q,q_{t+1};z\right)$ is L_t-Lipschitz with $L_t = \sqrt{3}\max\{D_q, L^C_{q,t+1}, L^C_{z,t}\}$.

To interpret the scaling of the result, we can appeal to van de Vaart and Wellner [1996, Thm. 2.6.4] which upper bounds the (log) covering numbers by the VC-dimension. Namely, van de Vaart and Wellner [1996, Thm. 2.6.4] states that there exists a universal constant K such that

$$N\left(\epsilon,\mathcal{F},L_r(Q)\right) \le KV(\mathcal{F})(4e)^{V(\mathcal{F})}\left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{F})-1)}.$$

Therefore, achieving an $\epsilon = cn^{-1}$ approximation error on the bracketing numbers of robust Q functions results in an $\log(2L_t n)$ dependence.

Lastly we remark on instantiating L_t. Note that under the assumption of bounded rewards, $D_{q,t+1} = B_r(T-t+1)$. Focusing on leading-order dependence in problem-dependent constants, we have that $L_t = O(B_r(T-t)\Lambda)$. Then $\hat{\mathcal{E}}(\hat{Q}) \le \epsilon + \sum_{t=1}^T K\frac{\log(2B_r(T-t)\Lambda n)}{n}$. Upper bounding the left Riemann sum by the integral, we obtain that

$$\sum_{t=1}^T K\frac{\log(2KB_r(T-t)\Lambda n/\epsilon)}{n} \le \int_1^T K\frac{\log(2KB_r(T-x)\Lambda n/\epsilon)}{n}dx = \frac{(T-1)}{n}(\log(2KB_r\Lambda(T-1)n/\epsilon)-1).$$

□

D.3 Confounding with infinite data

First, we prove the following useful result for confounded regression with conditional Gaussian tails:

Lemma 11. *Define:*

$$C(\Lambda) := \left(\frac{\Lambda^2 - 1}{\Lambda}\right) \phi\left(\Phi^{-1}\left(\frac{1}{1+\Lambda}\right)\right),$$

where ϕ and Φ are the standard Gaussian density and CDF respectively. Let $Y_t(Q)$ be conditionally Gaussian given $S_t = s$ and $A_t = a$ with mean $\mu_t(s,a)$ and standard deviation $\sigma_t(s,a)$. Then,

$$(\bar{\mathcal{T}}_t^* Q)(s,a) = \mu_t(s,a) - [1 - \pi_t^b(a|s)] C(\Lambda) \sigma_t(s,a).$$

Proof. Proof of Lemma 11 The CVaR for Gaussians has a closed-form [Norton et al., 2021]:

$$\frac{1}{1-\tau}\mathbb{E}_{\pi^b}\left[Y_t(Q)\mathbb{I}\left[Y_t(Q) < Z_t^{1-\tau}\right] | S_t = s, A_t = a\right] = \mu_t(s,a) - \sigma_t(s,a)\frac{\phi(\Phi^{-1}(1-\tau))}{1-\tau}.$$

Applying this to Proposition 5 gives the desired result. □

Proof. Proof of Proposition 7 First, note that R_t is conditionally Gaussian given S_t and A_t with mean $\theta_R\theta_P s$ and standard deviation $\theta_R\sigma_T$. Define $\beta_i := \theta_R \sum_{k=1}^{i} \theta_P^k$. Using value iteration, we can show that $V_{T-i}^{\pi^e}(s) = \beta_i s$ for $i \geq 1$. E.g. by induction, $V_{T-1}^{\pi^e}(s) = \theta_R\theta_P s = \beta_1$ and if $V_{T-t+1}^{\pi^e}(s) = \beta_{t-1}s$, then

$$V_{T-t}^{\pi^e}(s) = \theta_P(\theta_R + \gamma\beta_{t-1})s = \beta_t s.$$

Next we will derive the form of the robust value function by induction. For the base case, $t = T-1$, we have:

$$Y_{T-1} = \theta_R s'.$$

Therefore, Y_{T-1} is conditionally gaussian with mean $\theta_R\theta_P s$ and standard deviation $\theta_R\sigma_P$. Applying Lemma 11, we have:

$$\bar{V}_{T-1}^{\pi^e}(s) = \theta_R\theta_P s - 0.5C(\Lambda)\theta_R\sigma_P.$$

Now assume that $\bar{V}_{t+1}^{\pi^e}(s) = \theta_V s + \alpha_V$. Then

$$\begin{aligned} Y_t &= \theta_R s' + (\theta_V s' + \alpha_V) \\ &= (\theta_R + \theta_V)s' + \alpha_V. \end{aligned}$$

Therefore, Y_t is conditionally gaussian with mean $(\theta_R + \theta_V)\theta_P s + \alpha_V$ and standard deviation $(\theta_R + \theta_V)\sigma_P$. Applying Lemma 11, we have:

$$\bar{V}_t^{\pi^e}(s) = (\theta_R + \theta_V)\theta_P s + \alpha_V - 0.5C(\Lambda)(\theta_R + \theta_V)\sigma_P, \tag{31}$$

which is linear in s with new coefficients $\theta_V' := (\theta_R + \theta_V)\theta_T$ and $\alpha_V' := \alpha_V - 0.5C(\Lambda)(\theta_R + \theta_V)\sigma_P$.

By rolling out the recursion defined in Equation (31), consolidating the coefficients into β_i terms, and then simplifying we get:

$$\bar{V}_0^{\pi^e}(s) = V_0^{\pi^e}(s) - \frac{1}{2\theta_P}\left(\sum_{i=0}^{T-1}\beta_i\right)\sigma_P C(\Lambda).$$

Finally, that $C(\Lambda) \leq \frac{1}{8}\log(\Lambda)$ can be verified numerically.

□

D.4 Proofs for warm-starting

Proof of Theorem 2. We prove this via backwards induction.

Next we show asymptotic linearity. Define the following:

$$\theta^*_{t,a} = \mathbb{E}[\phi_{t,a}\phi_{t,a}^\top]^{-1}\mathbb{E}[\phi_{t,a}^\top \overline{Q}_t(S_t,a)] = \mathbb{E}[\phi_{t,a}\phi_{t,a}^\top]^{-1}\mathbb{E}[\phi_{t,a}^\top \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}^*_{t+1,a})]$$

$$\tilde{\theta}_{t,a} = \mathbb{E}_n[\phi_{t,a}\phi_{t,a}^\top]^{-1}\mathbb{E}_n[\phi_{t,a}^\top \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}^{(k)}_{t+1,a})] = \mathbb{E}_n[\phi_{t,a}\phi_{t,a}^\top]^{-1}\sum_{k=1}^K \mathbb{E}_k[\phi_{t,a}^\top \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}^{(k)}_{t+1,a})]$$

$$\hat{\theta}_{t,a} = \mathbb{E}_n[\phi_{t,a}\phi_{t,a}^\top]^{-1}\sum_{k=1}^K \mathbb{E}_k[\phi_{t,a}^\top \tilde{Y}_{t,a}(\zeta_t^{(k)}, \overline{\theta}^{(k)}_{t+1,a})]$$

Note that

$$\sqrt{n}(\hat{\theta}_{t,a} - \theta^*_{t,a}) = \sqrt{n}(\hat{\theta}_{t,a} - \tilde{\theta}_{t,a}) + \sqrt{n}(\tilde{\theta}_{t,a} - \theta^*_{t,a})$$

Orthogonality and cross-fitting in Proposition 6 establish that the first term is $o_p(1)$. The second term includes $\overline{\theta}^{(k)}_{t+1,a}$ as a generated regressor term, and we establish its asymptotic variance by GMM.

Note that

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_{t,a} - \tilde{\theta}_{t,a}) &= \mathbb{E}_n[\phi_{t,a}\phi_{t,a}^\top]^{-1}\sum_{k=1}^K \left\{ \mathbb{E}_k[\phi_{t,a}^\top \tilde{Y}_{t,a}(\hat{\zeta}_t^{(k)}, \hat{\overline{\theta}}^{(k)}_{t+1,a})] - \mathbb{E}_k[\phi_{t,a}^\top \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}^*_{t+1,a})] \right\} \\
&= \mathbb{E}_n[\phi_{t,a}\phi_{t,a}^\top]^{-1}\sum_{k=1}^K \left\{ \mathbb{E}\left[\phi_{t,a}^\top \left(\tilde{Y}_{t,a}(\hat{\zeta}_t^{(k)}, \hat{\overline{\theta}}^{(k)}_{t+1,a}) - \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}^*_{t+1,a}) \right) \right] \right\} \\
&\quad + \mathbb{E}_n[\phi_{t,a}\phi_{t,a}^\top]^{-1}\sum_{k=1}^K \left\{ (\mathbb{E}_k - \mathbb{E}) \left[\phi_{t,a}^\top \tilde{Y}_{t,a}(\hat{\zeta}_t^{(k)}, \hat{\overline{\theta}}^{(k)}_{t+1,a}) - \phi_{t,a}^\top \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}^*_{t+1,a}) \right] \right\}
\end{aligned}$$

We will show the first term is $o_p(n^{-\frac{1}{2}})$ by orthogonality. Define

$$S_{1,k} := \mathbb{E}\left[\phi_{t,a}^\top \left(\tilde{Y}_{t,a}(\hat{\zeta}_t^{(k)}, \hat{\overline{\theta}}^{(k)}_{t+1,a}) - \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}^*_{t+1,a}) \right) \right]$$

We consider elements of the vector-valued moment condition: for each $j = 1, \ldots, p$:

$$\begin{aligned}
&= \mathbb{E}\left[(\phi_{t,a}^\top)_j \mathbb{E}\left[\tilde{Y}_{t,a}(\hat{\zeta}_t^{(k)}, \hat{\overline{\theta}}^{(k)}_{t+1,a}) - \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}^*_{t+1,a}) \mid S_t, a \right] \right] \\
&\leq \|(\phi_{t,a}^\top)_j\| \|\mathbb{E}[\tilde{Y}_{t,a}(\hat{\zeta}_t^{(k)}, \hat{\overline{\theta}}^{(k)}_{t+1,a}) - \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}^*_{t+1,a}) \mid S_t, a]\| \\
&\leq \overline{C} \|\mathbb{E}[\tilde{Y}_{t,a}(\hat{\zeta}_t^{(k)}, \hat{\overline{\theta}}^{(k)}_{t+1,a}) - \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}^*_{t+1,a}) \mid S_t, a]\| && \text{by Assumption 11} \\
&= o_p(n^{-\frac{1}{2}}) && \text{by Proposition 6}
\end{aligned}$$

Next we study the sampling/cross-fitting terms:

$$S_{2,k} := \left\{ (\mathbb{E}_k - \mathbb{E}) \left[\phi_{t,a}^\top \left(\tilde{Y}_{t,a}(\hat{\zeta}_t^{(k)}, \hat{\overline{\theta}}^{(k)}_{t+1,a}) - \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}^*_{t+1,a}) \right) \right] \right\}$$

Since $|\mathcal{I}_k| \simeq n/K$, by the concentration of iid terms, by Cauchy-Schwarz inequality, we have that

$$S_{2,k} = o_p\left(n^{-1/2} \sum_{i=1}^{p} \mathbb{E}\left[\left(\tilde{Y}_{t,a}(\hat{\zeta}_t^{(k)}, \hat{\overline{\theta}}_{t+1,a}^{(k)}) - \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}_{t+1,a}^*)\right)^2 ((\phi_{t,a})_j)^2\right]^{1/2}\right)$$

Further,

$$\begin{aligned}
&\sum_{i=1}^{p} \mathbb{E}\left[\left(\tilde{Y}_{t,a}(\hat{\zeta}_t^{(k)}, \hat{\overline{\theta}}_{t+1,a}^{(k)}) - \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}_{t+1,a}^*)\right)^2 ((\phi_{t,a})_j)^2\right]^{1/2} \\
&\leq \overline{C}\|\tilde{Y}_{t,a}(\hat{\zeta}_t^{(k)}, \hat{\overline{\theta}}_{t+1,a}^{(k)}) - \tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}_{t+1,a}^*)\| \\
&= o_p(1) && \text{by consistency of nuisances}
\end{aligned}$$

Therefore, by continuous mapping theorem, Slutsky's theorem, and Assumption 11, $\sqrt{n}(\hat{\theta}_{t,a} - \tilde{\theta}_{t,a}) = o_p(1)$.

Next we study $\sqrt{n}(\tilde{\theta}_{t,a} - \theta_{t,a}^*)$. One approach for establishing asymptotic variance under generated regressors is via GMM, which we do so in this setting [Newey and McFadden, 1994]. We can write $\overline{\theta}$ as the parameter vector satisfying the "stacked" moment conditions (over timesteps and actions) at the true quantile parameter ζ (via our previous orthogonality analysis).

The moment functions for the robust Q-function parameters of interest, $\overline{\theta}_{t,\cdot}$, satisfy:

$$\left\{0 = \mathbb{E}\left[\left\{\tilde{Y}_{t,a}(\zeta_t^*, \overline{\theta}_{t+1}) - \overline{\theta}_{t,a}^\top \phi_{t,a}\right\} \phi_{t,a}^\top \mid A = a\right]\right\}_{a \in \mathcal{A}, t=1,\ldots,T} \tag{32}$$

We let these stacked moments be denoted as $\{0 = \mathbb{E}[g_{t,a}(\zeta^*, \bar{\theta})]\}_{a \in \mathcal{A}, t=0,\ldots,T-1}$.

For GMM, the asymptotic covariance matrix is given by

$$\sqrt{n}(\tilde{\theta}_{t,a} - \theta_{t,a}^*) \xrightarrow{d} -\left(G'G\right)^{-1} G' N(0, I) = N(0, V)$$

where $G = \partial g(\zeta^*, \bar{\theta})/\partial\theta$ and a consistent estimator of the asymptotic variance is given by $\hat{V} = \left(\hat{G}'\hat{G}\right)^{-1}, \hat{G} = \partial \hat{g}(\zeta^*, \bar{\theta})/\partial\bar{\theta}$.

Note that G is a block upper triangular matrix. The (blockwise) entries on the time diagonal are given by the covariance matrix $\phi_{t,a}\phi_{t,a}^\top$ (i.e., from linear regression). The lower entries, i.e. $\partial g_{t,a}(\zeta^*, \overline{\theta})/\partial\overline{\theta}_{t+1,a'}$ are given below, by differentiating under the integral:

$$\begin{aligned}
&\frac{\partial g_{t,a}(\zeta^*, \overline{\theta})}{\partial\overline{\theta}_{t+1,a'}} \Big\{\mathbb{E}\Big[\Big\{\Big(\alpha_{t,a} Y_{t,a}(\overline{\theta}_{t+1}) + (1 - \alpha_{t,a}) \cdot \tfrac{1}{1-\tau}\Big(Y_{t,a}(\overline{\theta}_{t+1})\mathbb{I}(Y_{t,a}(\overline{\theta}_{t+1}) \leq \zeta_{t,a}^\top \phi_{t,a}) \\
&\qquad -\zeta_{t,a}^\top \phi_{t,a} \cdot [\mathbb{I}\{Y_{t,a}(\overline{\theta}_{t+1}) \leq \zeta_{t,a}^\top \phi_{t,a}\} - (1-\tau)]\Big)\Big) - \overline{\theta}_{t,a}{}^\top \phi_{t,a}\Big\} \phi_{t,a}^\top \mid A = a\Big]\Big\} \\
&= \mathbb{E}\left[\alpha_{t,a}(\phi(S_{t+1}, a')\phi(S_t, A_t)^\top) + \frac{1-\alpha_{t,a}}{1-\tau}\left(\int_{-\infty}^{q} (\phi(S_{t+1}, a_{t+1})) dP_{S_{t+1}|S_t, A_t}\right) \phi^\top(S_t, A_t) \mid A_t = a\right] \\
&= \mathbb{E}[\alpha_{t,a}(\phi(S_{t+1}, a')\phi(S_t, a)^\top)] + \mathbb{E}\left[(1-\alpha_{t,a})(\mathbb{E}[\phi(S_{t+1}, a_{t+1}) \mid Y_{t+1} \leq \zeta_{t,a}^\top \phi_{t,a}, S_t, A_t = a]\phi^\top(S_t, a)\right]
\end{aligned}$$

Denote $Z_{a_{t+1}}^{\phi_{t+1}}(S_t, a) = \mathbb{E}[\phi(S_{t+1}, a_{t+1}) \mid Y_{t+1} \leq \zeta_{t,a}^\top \phi_{t,a}, S_t, A_t = a]$, then

$$
\begin{aligned}
\frac{\partial g_{t,a}(\zeta^*, \overline{\theta})}{\partial \bar{\theta}_{t+1,a'}} &= \mathbb{E}\left[\alpha_{t,a}(\phi(S_{t+1}, a')\phi(S_t, a)^\top) + (1-\alpha_{t,a})(Z^{\phi}_{a'}(S_t, a)\phi^\top(S_t, a))\right] \\
&= \mathbb{E}\left[\alpha_{t,a}(\phi_{t+1,a'}\phi_{t,a}^\top) + (1-\alpha_{t,a})(Z^{\phi}_{a',t,a}\phi_{t,a}^\top)\right] \\
&= \tilde{\Sigma}_{t,a}^{t+1,a'} + \tilde{\Omega}_{t,a}^{a',t,a}
\end{aligned}
$$

So, G is a block upper triangular matrix:

$$
\begin{bmatrix}
\ddots & \dots & & & \\
0 & \mathbb{E}[\phi_{t,a}\phi_{t,a}^\top] & \{\tilde{\Sigma}_{t,a}^{t+1,a'} + \tilde{\Omega}_{t,a}^{a',t,a}\}_{a'\in\mathcal{A}} & & \\
0 & 0 & \ddots & \{\tilde{\Sigma}_{t,a_k}^{t+1,a'} + \tilde{\Omega}_{t,a_k}^{a',t,a_k}\}_{a'\in\mathcal{A}} & \\
0 & 0 & 0 & \mathbb{E}[\phi_{t,a_K}\phi_{t,a_K}^\top] & \{\tilde{\Sigma}_{t,a_K}^{t+1,a'} + \tilde{\Omega}_{t,a_K}^{a',t,a_K}\}_{a'\in\mathcal{A}}
\end{bmatrix}
$$

$\square$

E Details on experiments

E.1 Parameter Values

$\theta_A = -0.05$, $\sigma = 0.36$, $\gamma = 0.9$, $H = 4$.

The matrices A and B were chosen randomly with a fixed random seed:

```
np.random.seed(1)
B_sparse0 = np.random.binomial(1,0.3,size=d)
B = 2.2*B_sparse0 * np.array( [ [ 1/(j+k+1) for j in range(d) ]
                                            for k in range(d) ] )

np.random.seed(2)
A_sparse0 = np.random.binomial(1,0.6,size=d)
A = 0.48*A_sparse0 * np.array( [ [ 1/(j+k+10) for j in range(d) ]
                                              for k in range(d) ] )
```

Likewise for θ_R:

```
theta_R = 3 * np.random.normal(size=d) * np.random.binomial(1,0.3,size=d)
```

E.2 Function Approximation

Conditional expectations were approximated with the Lasso using `scikit-learn`'s implementation, with regularization hyperparameter α = 1e-4. Conditional quantiles were approximated with `scikit-learn`'s ℓ_1-penalized quantile regression, regularization hyperparameter *alpha* = 1e-2, using the `highs` solver.

E.3 Calculating Ground Truth

To provide ground truth for our sparse linear setting, we analytically derive the form of the robust Bellman operator. Consider the candidate Q function, $Q(s,0) = \beta^T s + a_0, Q(s,1) = \beta^T + a_1$. Then,

$$\begin{aligned} Y_t &= \theta_R^T S_{t+1} + \gamma \beta^T S_{t+1} + \theta_A \gamma \max\{1_d^T \theta_R, 0\} \\ &= \theta_R^T S_{t+1} + \gamma \beta^T S_{t+1} + \theta_A \gamma 1_d^T \theta_R \end{aligned}$$

where we chose simulation parameters such that $\theta_A \gamma \max\{1_d^T \theta_R, 0\} > 0$. Therefore:

$$Y_t | S_t, A_t \sim \mathcal{N}\left((\theta_R + \gamma\beta)^T (B S_t + \theta_A A_t) + \theta_A \gamma 1_d^T \theta_R, \sqrt{\sum_{i=1}^d (\theta_R + \gamma\beta)_i^2 (A S_t + \sigma)_i^2} \right)$$

Since Y_t is conditionally Gaussian, we apply Lemma 11:

$$\begin{aligned} (\bar{\mathcal{T}}_t^* Q)(s,a) &= \mathbb{E}[Y_t | S_t = s, A_t = a] - 0.5 C(\Lambda) \sqrt{\mathrm{Var}[Y_t | S_t = s, A_t = a]} \\ &= (\theta_R + \gamma\beta)^T (B S_t + \theta_A A_t) + \theta_A \gamma 1_d^T \theta_R - 0.5 C(\Lambda) \sqrt{\sum_{i=1}^d (\theta_R + \gamma\beta)_i^2 (A S_t + \sigma)_i^2} \end{aligned}$$

First, note that the slope w.r.t. S_t is not a function of A_t validating our choice of an action-independent β. Second, note that only the last term is non-linear in S_t. So the ground truth for FQI with Lasso adds the first two terms to the closest linear approximation of this last term. Since our object of interest is the average optimal value function at the initial state, we perform this linear approximation in terms of mean squared error at the initial state. In practice, we compute this by drawing $200,000$ samples i.i.d. from the initial state distribution and then doing linear regression on this last term. Plugging the slope and intercept back in is extremely close to the best linear approximation of $(\bar{\mathcal{T}}_t^* Q)(s,a)$.

F High-Dimensional Experiments

The baseline experiments in the main text used $d = 25$ features. A and B are chosen to be column-wise sparse, with 5 and 20 non-zero columns respectively. The sample size was $n = 5000$ so with a uniform behavior policy, we have approximately 2500 samples of each action. Therefore, for each Lasso regression, the sample size is 100 times larger than the number of dimensions — a fairly low dimensional setting. Here, we repeat this experiment in a higher-dimensional setting with $d = 100$, and $n = 600$. Again with a uniform behavior policy, this means around 300 samples for 100 features in each Lasso regression, a much higher dimensional problem. As before, A and B are chosen to be column-wise sparse, with roughly 2/3rds and 1/3rds of the columns non-zeros respectively (the same fraction as in the main paper).

F.1 Parameter Values

$\theta_A = -0.05$, $\sigma = 0.1$, $\gamma = 0.9$, $H = 4$.

The matrices A and B were chosen randomly with a fixed random seed:

```
np.random.seed(1)
B_sparse0 = np.random.binomial(1,0.3,size=d)
B = 2.2*B_sparse0 * np.array( [ [ 1/(j+k+1) for j in range(d) ]
                                            for k in range(d) ] )/1.2

np.random.seed(2)
A_sparse0 = np.random.binomial(1,0.6,size=d)
A = 0.48*A_sparse0 * np.array( [ [ 1/(j+k+10) for j in range(d) ]
                                              for k in range(d) ] )/20
```

Likewise for θ_R:

```
theta_R = 2 * np.random.normal(size=d) * np.random.binomial(1,0.3,size=d)
```

F.2 Results

We repeat the same metrics as in the main text. See Table 2 for the results.

Notice that in this setting, policy optimization is *substantially* harder — even the nominal policy estimate only picks the true optimal action 28% of the time. However, we still see almost identical behavior as in the low-dimensional setting when comparing the orthogonal and non-orthogonal. Without orthogonalization, performance drops off dramatically as Λ increases. Our orthogonalized algorithm has MSE that decays more gracefully with Λ, and picks the correct action at essentially the same rate as the nominal algorithm.