Efficient and Sharp Off-Policy Evaluation in Robust Markov Decision Processes

Andrew Bennett[*]
Morgan Stanley
andrew.bennett@morganstanley.com

Nathan Kallus[*]
Cornell University
kallus@cornell.edu

Miruna Oprescu[*]
Cornell University
amo78@cornell.edu

Wen Sun[*]
Cornell University
ws455@cornell.edu

Kaiwen Wang[*]
Cornell University
kw437@cornell.edu

Abstract

We study evaluating a policy under best- and worst-case perturbations to a Markov decision process (MDP), given transition observations from the original MDP, whether under the same or different policy. This is an important problem when there is the possibility of a shift between historical and future environments, due to *e.g.* unmeasured confounding, distributional shift, or an adversarial environment. We propose a perturbation model that can modify transition kernel densities up to a given multiplicative factor or its reciprocal, which extends the classic marginal sensitivity model (MSM) for single time step decision making to infinite-horizon RL. We characterize the sharp bounds on policy value under this model, that is, the tightest possible bounds given by the transition observations from the original MDP, and we study the estimation of these bounds from such transition observations. We develop an estimator with several appealing guarantees: it is semiparametrically efficient, and remains so even when certain necessary nuisance functions such as worst-case Q-functions are estimated at slow nonparametric rates. It is also asymptotically normal, enabling easy statistical inference using Wald confidence intervals. In addition, when certain nuisances are estimated inconsistently we still estimate valid, albeit possibly not sharp bounds on the policy value. We validate these properties in numerical simulations. The combination of accounting for environment shifts from train to test (robustness), being insensitive to nuisance-function estimation (orthogonality), and accounting for having only finite samples to learn from (inference) together leads to credible and reliable policy evaluation.

1 Introduction

Offline policy evaluation (OPE) from historical data is crucial in domains where active, on-policy experimentation is costly, risky, unethical, or otherwise operationally infeasible. Relevant domains range from medicine, to finance, to recommendation systems. However, whenever historical data is used to study future behavior, there is a concern of non-stationarity – shift between the environment generating the data (training environment) and the environment in which a policy will be deployed (test environment). This may occur, *e.g.*, due to general distributional shifts in the environment over time, unobserved confounding in the observed historical data, or adversarial elements of the environment (such as other agents) that may react when the agent is deployed. So, even as standard OPE in offline reinforcement learning (ORL) accounts for the change between the logging and evaluation policies, it may fail to account for the fact that the Markov decision process (MDP) too

[*] Alphabetical order.

Preprint. Under review.

has changed. While this issue is particularly critical in high-stakes domains, it is broadly appealing to understand how value shifts across different environments in any application domain.

Robust MDPs [30, 48] model unknown environments by allowing an adversary to choose from any one environment in a set. Therefore, they offer a natural model for unknown environment shifts by simply considering all environments to which we could possibly shift. A variety of work addresses questions such as planning in a known robust MDP [27, 44, 65] as well as online learning [5, 64]. Here we focus on a purely statistical estimation question: given observations of transitions from some unknown transition kernel, we wish to estimate the worst-case (or best-case) value of a given evaluation policy in a robust MDP, defined by a set of MDPs whose transition functions are centered around the observed transition kernel. This setting captures the previously studied unconfounded robust OPE problem [63], where the observed transition kernel corresponds to an MDP, and the observed transitions are the result of applying some logging policy within this MDP, in which case we seek policy value estimates that are robust to future changes in the MDP dynamics. However, our setting is more general than this, and also captures problems where the observed transitions are confounded by some unobserved variables, in which case they do *not* correspond to observations from the transition kernel of an MDP. In this case, the robust MDP and the robust policy value estimates are designed to account for worst-case (or best-case) impact of this confounding bias. In either case, as in ORL, we emphasize that we do *not* know the observational MDP, and can only access it via a samples of transitions. Furthermore, even in the simple case with no unmeasured confounding, in a notable departure from standard ORL, the problem can be difficult even if the logging and evaluation policies are the same (the usually easy on-policy setting), since the policy can induce very different visitation distributions in the original and perturbed MDPs.

Such robust offline evaluation from transition data was considered in recent work [11, 50]. We build on this recent work by focusing the question of statistically *efficient* and *robust* estimation of the *sharp* bounds (*i.e.*, the tightest possible given the data). Previous work focused on evaluation using only the Q-function under the worst-case environment (in some cases under a relaxation of the adversary, leading to loose bounds). Thus, any error in its estimation translates directly to error in evaluation. Thus flexible nonparametric modeling of this function can mean slow rates for estimated bounds and a lack semiparametric efficiency. Moreover, lacking an understanding of the noise in the estimates, we cannot add confidence bands on top of the bounds, leading to the risk of bounds that are too small.

We address these issues by developing an orthogonalized estimation method. Our approach combines several nuisance functions: the Q-function in the worst-case environment, the state-visitation frequency in the worst-case environment, and a threshold characterizing the worst-case environment. Our first key result is that our estimator is to first orders as though we took a sample average with the true value of these functions without having to estimate them at all, provided we just estimate them at certain slow nonparametric rates. This ensures we not only have a $\sqrt{n}$-rate of estimation even when nuisances are estimated more slowly, but also that we are asymptotically normal, permitting the construction of confidence bands on top of the bounds so that we are assured an actual bound with high confidence. We further show that our asymptotic variance is in fact the minimum achievable, that is, we attain semiparametric efficiency. Our second key result show that even if we do not estimate some of the nuisance functions correctly, we are still consistent to sharp or valid bounds (depending on the case). That is, even when we are biased due to misestimation of nuisances, our bias (if any) only enlarges our bounds, so they remain valid. We illustrate these guarantees numerically. Taken together, these guarantees provide considerable credibility to bounds produced by our method.

Our contributions are summarized as follows:

1. We provide new algorithms and analysis for learning robust Q-functions (Section 3) and robust visitation density ratios (Section 4) under the function approximation setting.
2. We derive the sharp and efficient estimator for the robust policy value, which is optimal in the local-minimax sense and is the gold standard in semiparametric estimation (Section 5).
3. We empirically validate the efficiency and sharpness of our approach (Section 6).

1.1 Related Works

Unobserved Confounding in Sequential Decision Making. OPE in robust MDPs is related to OPE bounds in confounded MDPs, where the behavior policy and the transition kernel are influenced by unobserved confounders. The constraint Eq. (1) defining our target robust MDP aligns with the

Marginal Sensitivity Model (MSM) [56] employed in sensitivity analysis for causal inference. Yet, unlike the MSM, which limits the ratio of policy densities, our approach directly constrains the ratio of the transition kernels. Our formulation can be viewed as a generalization of the MSM from traditional two-action no-horizon causal effects (where the constrains coincide) to multi-action infinite-horizon discounted MDPs, where the next state is the "potential outcome." In that sense, our model essentially serves as an outcome-based sensitivity model [9]. This distinction is crucial as it enables our model to subsume the policy-based MSM in cases where the policy is confounded. Nonetheless, the reverse does not hold, and the policy-based MSM does not imply a transition kernel-based MSM for $A > 2$. This point is further corroborated by [11], who explore the policy-based MSM within confounded MDPs and obtain *non-sharp* identification bounds when $A > 2$. In contrast, our approach yields *sharp* identification in general, regardless of the number of actions and without placing assumptions on the behavior policy, which may or may not be confounded.

[12] also considered an MSM-like model in the transition kernel but their formulation assumes $A = 2$. [35] operates under the setting of [11] and required tabular states. We note that all these works including ours considers *i.i.d.* confounders at each step, which translates to a robust MDP with (s, a)-rectangularity and ensures that the worst-case problem is still an MDP rather than a POMDP. The importance of this assumption was verified by [47], who showed that without it the non-memoryless confounder can create exponential-in-horizon changes in value.

Neyman Orthogonality and Semiparametric Efficient Estimation. We leverage a body of research focusing on learning with nuisances functions (e.g., Q-functions) that we need to estimate from data but are not the primary target (e.g., policy value). Much of this research [6, 15, 16, 26, 54, 60, among others] aims to identify Neyman-orthogonal estimators, which are first order orthogonal (insensitive) to nuisance errors. This literature is tightly linked to the semiparametric efficient estimation literature since Neyman-orthogonal scores can arise naturally from efficient influence functions [29, 53]. Going beyond the no-horizon causal inference setting, some explore such estimators in off-policy sequential-decisions contexts [18, 33, 36, 41, 43]. Notably, [34] derive efficient influence functions and orthogonal estimation in standard, non-robust OPE in infinite-horizon RL, which coincides with our unconfounded no-uncertainty case ($\Lambda = 1$).

Going beyond point-identified settings, some explore orthogonality and efficiency for partial identification and sensitivity analysis. In the causal inference literature, efficient/orthogonal estimation in the no-horizon setting has been studied extensively under several sensitivity models [9, 17, 21]. Closest to our work is [21] who provide an orthogonal estimator and convergence rates under the MSM [56], which coincides with our setting under $\gamma = 0$. In the sequential setting, [47] considers confounding at a single time step under the MSM, but their estimator is not orthogonal when the quantile function is unknown. [11] provide a fitted-Q-iteration learner with an orthogonalized loss function, but not orthogonal/efficient estimates of worst-case policy value.

2 Preliminaries

We consider an MDP with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $P(s' \mid s, a)$, reward function $r(s, a) \in [0, 1]$ and initial state distribution $d_1 \in \Delta(\mathcal{S})$. We do not require $\mathcal{S}$ or $\mathcal{A}$ to be finite. We assume r and d_1 are known for simplicity, and it is standard to extend our analysis to when they are unknown. We are given a dataset $\mathcal{D}$ of n *i.i.d.* tuples (s_i, a_i, r_i, s_i') such that $(s_i, a_i) \sim \nu$, $s_i' \sim P(\cdot \mid s, a)$ and $r_i = r(s_i, a_i)$, where ν is an arbitrary data-generating distribution. For discount factor $\gamma \in [0, 1)$, let the Q function be the discounted cumulative rewards under a policy $\pi : \mathcal{S} \to \mathcal{A}$, $Q_{\pi,P}(s, a) = \mathbb{E}_{\pi,P}[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t) \mid s_1 = s, a_1 = a]$. Similarly, define the value function as $V_{\pi,P}(s) = Q_{\pi,P}(s, \pi)$, where $f(s, \pi) := \mathbb{E}_{a \sim \pi(s)}[f(s, a)]$.

We are interested in estimating the value of a fixed target policy π_{t} (a.k.a. evaluation policy) in an unobserved MDP with a feasible perturbed transition kernel U. We say U is a feasible perturbation of the observed, nominal kernel P if for all s, a, s': we have

$$\Lambda^{-1}(s, a) \leq \frac{\mathrm{d}U(s' \mid s, a)}{\mathrm{d}P(s' \mid s, a)} \leq \Lambda(s, a) \tag{1}$$

where $\Lambda(s, a) \in [1, \infty)$ is a sensitivity parameter chosen by the practitioner. We denote the set of all feasible perturbations of P by $\mathcal{U}(P)$, which is an s, a-rectangular set [44]. We define the best- and worst-case Q functions of π_{t} as

$$Q^+(s, a) := \sup\nolimits_{U \in \mathcal{U}(P)} Q_{\pi_{\mathrm{t}},U}(s, a); \qquad Q^-(s, a) := \inf\nolimits_{U \in \mathcal{U}(P)} Q_{\pi_{\mathrm{t}},U}(s, a). \tag{2}$$

Thus, the goal of this paper is to estimate the best- and worst-case value of π_{t} at the initial state,

$$V_{d_1}^{\pm} := (1-\gamma)\mathbb{E}_{s_1 \sim d_1}[V^{\pm}(s_1)]. \tag{3}$$

where $V^{\pm}(s) = \mathbb{E}_{a\sim\pi_{\mathrm{t}}(s)}[Q^{\pm}(s,a)]$ and the $\pm$ symbol signals that an equation should be read twice, once with $\pm = +$ and once with $\pm = -$. This robust OPE problem (Eq. (3)) is much more challenging than standard OPE since the best- and worst-case transition kernels $U^{\pm}$ are unobserved as our dataset $\mathcal{D}$ is generated under P. For example, standard OPE is easy in the on-policy case *i.e.*, if $\mathcal{D}$ were generated by π_{t}, but our problem is still "off-data" and non-trivial.

Discounted Visitation Distributions. For any transition kernel U, define the discounted visitation distribution of π_{t} under U as: $d_{d_1,U}^{\pi_{\mathrm{t}},\infty}(s) := (1-\gamma)\sum_{h=1}^{\infty}\gamma^{h-1}d_{d_1,U}^{\pi_{\mathrm{t}},h}(s)$, where $d_{d_1,U}^{\pi_{\mathrm{t}},h}(s)$ is the probability of reaching state s in the Markov chain induced by transition kernel U and policy π_{t} starting from $d_1(\cdot)$. We will use $d^{\pm,\infty}$ as shorthand for $d_{d_1,U^{\pm}}^{\pi_{\mathrm{t}},\infty}$, where $U^{\pm}$ denotes the best- and worst-case kernel in $\mathcal{U}(P)$.

Bellman-type Operators. For any function $f : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ and transition kernel U, recall the Bellman operator is defined as $\mathcal{T}_U f(s,a) := r(s,a) + \gamma\mathbb{E}_U[f(s',\pi_{\mathrm{t}}) \mid s,a]$. For robust OPE, we define the following robust analogs $\mathcal{T}_{\mathsf{rob}}^{+} f(s,a) := r(s,a) + \gamma\sup_{U\in\mathcal{U}(P)}\mathbb{E}_U[f(s',\pi_{\mathrm{t}}) \mid s,a]$ and $\mathcal{T}_{\mathsf{rob}}^{-} f(s,a) := r(s,a) + \gamma\inf_{U\in\mathcal{U}(P)}\mathbb{E}_U[f(s',\pi_{\mathrm{t}}) \mid s,a]$. Moreover, we define $\mathcal{J}_U f(s,a) := \gamma\mathbb{E}_U[f(s',\pi_{\mathrm{t}}) \mid s,a] - f(s,a)$. For any linear operator $\mathcal{T}$, also let $\mathcal{T}'$ denote its adjoint: that is, for all $f,g\in L_2(\nu)$, $\langle f,\mathcal{T}g\rangle = \langle\mathcal{T}'f,g\rangle$, where $\langle\cdot,\cdot\rangle$ is the inner product in $L_2(\nu)$.

Conditional Value-at Risk (CVaR). For a random variable X, its upper/lower CVaRs at level $\tau\in[0,1]$ is defined as the average outcome of the upper/lower τ-fraction of cases, and are formally defined as follows [52]:

$$\mathrm{CVaR}_\tau^{+}(X) := \min\nolimits_{b\in\mathbb{R}}\{b+\tau^{-1}\mathbb{E}[(X-b)_+]\},\ \mathrm{CVaR}_\tau^{-}(X) := \max\nolimits_{b\in\mathbb{R}}\{b+\tau^{-1}\mathbb{E}[(X-b)_-]\},$$

where $y_+ := \max(0,y)$ and $y_- := \min(0,y)$ for $y\in\mathbb{R}$. The optima are attained at the upper/lower τ-th quantile of X which we denote as $\beta_\tau^{+}(X)/\beta_\tau^{-}(X)$, *i.e.*, $\mathrm{CVaR}_\tau^{\pm}(X) = \beta_\tau^{\pm}(X) + \tau^{-1}\mathbb{E}[(X-\beta_\tau^{\pm}(X))_{\pm}]$. If X has a cumulative distribution function (CDF) which is differentiable at $\beta_\tau^{\pm}(X)$, its CVaRs simplify to $\mathrm{CVaR}_\tau^{+}(X) = \mathbb{E}[X \mid X\geq\beta_\tau^{+}(X)]$ and $\mathrm{CVaR}_\tau^{-}(X) = \mathbb{E}[X \mid X\leq\beta_\tau^{-}(X)]$.

Notations. We use $x\lesssim y$ to mean that $x\leq Cy$ holds for some universal constant C. The indicator function $\mathbb{I}[p]$ takes value 1 if p is true and 0 otherwise. For a measure μ, we let $\|f\|_\mu := (\mathbb{E}_\mu|f(X)|^2)^{1/2}$ denote the L_2 norm of f, provided it exists. When μ is clear from context, we also use $\|f\|_p := (\mathbb{E}|f(X)|^p)^{1/p}$ to denote the L_p norm of f and $\|f\|_{p,n} := (\mathbb{E}_n|f(X)|^p)^{1/p}$ to denote the empirical analog. For a data sample of size n, we define the empirical mean as $\mathbb{E}_n[f(X)] = \frac{1}{n}\sum_{i=1}^{n}f(x_i)$. For a nuisance function f, we reserve f^* as its true value and $\widehat{f}$ as the learned value from data. Moreover, we employ $+$ and $-$ to denote functions corresponding to best- and worst-case bounds, respectively. The $\pm$ symbol signals that an equation should be read twice, once with $\pm = +$ and once with $\pm = -$. For example, $a_\pm + b_\pm = c_\pm$ is compact notation for two equations: $a_+ + b_+ = c_+$ and $a_- + b_- = c_-$. See Appendix A for a comprehensive notation table.

3 Robust Q-function and Estimation with Fitted-Q Evaluation

In this section, we identify the robust Q-function using the robust Bellman equation and then derive convergence rates for iteratively minimizing the robust Bellman error.

3.1 Identification of $Q^{\pm}$

The robust Q-functions of π_{t}, denoted as $Q^{\pm}$, satisfy the robust Bellman equations $Q^{\pm}(s,a) = \mathcal{T}_{\mathsf{rob}}^{\pm}Q^{\pm}(s,a), \forall s,a$ since the uncertainty set $\mathcal{U}(P)$ factorizes over s,a [30]. While these equations seem intractable due to the sup/inf in the definition of $\mathcal{T}_{\mathsf{rob}}^{\pm}$, [11] showed that $\mathcal{T}_{\mathsf{rob}}^{\pm}$ has a closed form solution in terms of CVaR under the *observed* kernel P.

Lemma 3.1. *Set* $\tau(s,a) = (\Lambda(s,a)+1)^{-1}$. *Then, for any* $q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$,

$$\mathcal{T}_{\mathsf{rob}}^{\pm}q(s,a) = r(s,a) + \gamma\Lambda^{-1}(s,a)\mathbb{E}[v(s') \mid s,a] + \gamma(1-\Lambda^{-1}(s,a))\,\mathrm{CVaR}_{\tau(s,a)}^{\pm}[v(s') \mid s,a],$$

where $v(s') = \mathbb{E}_{a'\sim\pi_t(s')}[q(s',a')]$, *and* $\mathbb{E}, \mathrm{CVaR}_\tau$ *are under the observed kernel* $P(\cdot \mid s,a)$.

Algorithm 1 RobustFQE: Iterative fitting for estimating $Q^{\pm}$ and $\beta_{\tau}^{\pm}$.

```
1: Input: Number of iterations $M$, Dataset $\mathcal{D}$ of size $n$, $Q$-function class $\mathcal{Q}$.
2: Initialize $\widehat{v}_0^{\pm}(s') = 0$.
3: for $i = 1, 2, \ldots, M$ do
4:     Create a sub-dataset $\mathcal{D}_i = \mathcal{D}[ni/M : n(i+1)/M]$.
5:     On the first half of $\mathcal{D}_i$, estimate the $\tau(s,a)$-quantiles of $\widehat{v}_{i-1}^{\pm}(s'), s' \sim P(\cdot \mid s,a)$. Let $\widehat{\beta}_i^{\pm}(s,a)$
       denote the learned upper/lower quantiles from the quantile regression oracle QR.
6:     Using the second half of $\mathcal{D}_i$, solve the empirical robust Bellman equation by minimizing
       squared prediction error for the pseudo-outcome:
           $\widehat{q}_i^{\pm} \leftarrow \arg\min_{q \in \mathcal{Q}} \widehat{\mathbb{E}}_{\mathcal{D}_i[|\mathcal{D}_i|/2+1:]}[(y^{\pm}(s,a,s') - q(s,a))^2]$,
       where $y^{\pm}(s,a,s') = r(s,a) + \gamma\Lambda^{-1}(s,a)\widehat{v}_{i-1}^{\pm}(s') + \gamma(1 - \Lambda^{-1}(s,a))$
           $\times (\widehat{\beta}_i^{\pm}(s,a) + \tau^{-1}(s,a)(\mathbb{E}_{a' \sim \pi_{\mathsf{t}}(s')}[\widehat{q}_i^{\pm}(s',a')] - \widehat{\beta}_i^{\pm}(s,a))_{\pm})$.
7: Output: $\widehat{q}_M^{\pm}, \widehat{\beta}_M^{\pm}$.
```

3.2 Estimating the robust Q-function with Robust FQE

We now show that $Q^{\pm}$ can be estimated via an iterative fitting algorithm inspired by the standard fitted Q-evaluation (FQE) [46]. Our algorithm RobustFQE (Algorithm 1) proceeds for M iterates with two main steps in each iterate. First, in Line 5, we estimate the upper- and lower-quantile of $\widehat{v}_{i-1}(s') \mid s,a$. Here, we assume access to an oracle QR for quantile regression oracle, which is a well-understood problem and many algorithms can be plugged in. Second, in Line 6, we solve the tractable robust Bellman equation in Lemma 3.1 with the CVaR term estimated by its orthogonal estimating equation with the learned quantiles [49]. By orthogonally estimating CVaR, we obtain second-order dependence on the quantile estimation errors of the first step. we minimize mean squared error with a general function class $\mathcal{Q} \subset \mathcal{S} \times \mathcal{A} \mapsto [0, (1-\gamma)^{-1}]$, which can be linear functions or neural networks.

We make two assumptions. First, we assume that the quantile regression oracle has a convergence rate, which is convergence rate quantile regression oracle has a convergence rate, which be guaranteed under smoothness conditions [8, 13, 24, 25, 45, 51, 55].

Assumption 3.2 (QR Oracle)**.** For any $v : \mathcal{S} \mapsto [0, (1-\gamma)^{-1}]$, let the true $\tau(s,a)$-quantile of $v(s'), s' \sim P(s,a)$ be denoted by $\beta_{\tau}^{v}(s,a)$. Given a dataset $\mathcal{D}_{\mathsf{QR}}$, we assume QR outputs estimates $\widehat{\beta}_v$ with bounded ℓ_∞ error: for any δ, w.p. $1-\delta$, $\|\widehat{\beta}_q - \beta_{\tau}^{q}\|_\infty < \mathsf{err}_{\mathsf{QR}}(|\mathcal{D}_{\mathsf{QR}}|, \delta)$.

The second assumption is completeness under the robust Bellman $\mathcal{T}_{\mathsf{rob}}^{\pm}$. Completeness is a standard assumption in model-free RL; without it, fitted-Q can diverge or converge to bad fixed points [38, 58].

Assumption 3.3 (Completeness)**.** For all $q \in \mathcal{Q}$, we have $\mathcal{T}_{\mathsf{rob}}^{\pm} q \in \mathcal{Q}$.

We note that the current proofs of [11, 50] require a stronger completeness: $\mathcal{T}_{\beta} q \in \mathcal{Q}$ for all $q \in \mathcal{Q}$ and feasible β. We circumvent the need for the stronger "all-β" completeness by bounding model misspecification of least squares regression with second order error in the quantile regression.

Finally, we express our bounds with the critical radius $\varepsilon_n^{\mathcal{Q}}$, a standard tool for deriving fast rates in statistics; see Appendix D for a summary. Also, we denote the standard concentrability coefficient with $C_{d_1}^{\pm} := \left\| \mathrm{d}d_{\mu}^{\pm,\infty} / \mathrm{d}d_1 \right\|_\infty$, a standard and necessary quantity for OPE.

Theorem 3.4. *Let $\varepsilon_n^{\mathcal{Q}}$ denote the critical radius of $\mathcal{Q}$. Under Assumptions 3.2 and 3.3, RobustFQE ensures that for any $\delta \in (0,1)$, w.p. $1-\delta$,*

$$\|\widehat{q}_M^{\pm} - Q^{\pm}\|_{d_1} \lesssim (1-\gamma)^{-2}\left(\sqrt{C_{d_1}^{\pm}} \cdot \varepsilon_n^{\mathcal{Q}} + \mathsf{err}_{\mathsf{QR}}^2(n/2M, \delta/2M)\right), \ \ \textit{and}$$

$$\left|(1-\gamma)\mathbb{E}_{d_1}[\widehat{v}_M^{\pm}(s_1)] - V_{d_1}^{\pm}\right| \lesssim \gamma^M + (1-\gamma)^{-1}\left(\sqrt{C_{d_1}^{\pm}} \cdot \varepsilon_n^{\mathcal{Q}} + \mathsf{err}_{\mathsf{QR}}^2(n/2M, \delta/2M)\right).$$

For parametric classes (*e.g.*, finite or linear), the critical radius converges at the standard $\widetilde{\mathcal{O}}(n^{-1/2})$ rate. Thanks to the orthogonal estimation of CVaR, we also have a favorable second-order dependence

Algorithm 2 RobustMIL: Minimax Estimation of $w^{\pm}$ with a Stabilizer

```
Input: Dataset $\mathcal{D}$, prior stage estimate $\widetilde{\zeta}$, function classes $\mathcal{W}, \mathcal{F}$, stabilizer weight $\lambda > 0$.
Define weights $\xi^{\pm}(s,a,s') := \Lambda^{-1}(s,a) + (1-\Lambda^{-1}(s,a))\tau^{-1}(s,a)\mathbb{I}[\pm\widetilde{\zeta}(s,a,s') \geq 0]$.
Output:
    $\widehat{w}^{\pm} = \arg\min_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} \mathbb{E}_n\left[w(s,a)(\gamma\xi^{\pm}(s,a,s')f(s',\pi_{\mathrm{t}}) - f(s,a)) + (1-\gamma)\mathbb{E}_{d_1} f(s_1,\pi_{\mathrm{t}})\right] - \lambda\|\gamma\xi^{\pm}(s,a,s';\widetilde{\zeta})f(s',\pi_{\mathrm{t}}) - f(s,a)\|_{2,n}^2 \quad (6)$
```

on $\mathrm{err}_{\mathsf{QR}}$ which allows for quantile regression to converge at slower $\widetilde{\mathcal{O}}(n^{-1/4})$ rates. The main disadvantage of this direct approach is that it converges at a slow sub-$\sqrt{n}$ rate if $\varepsilon_n^{\mathcal{Q}}$ converges at a sub-$\sqrt{n}$, *e.g.*, $\varepsilon_n^{\mathcal{Q}}$ converges at a $\widetilde{\mathcal{O}}(n^{-1/4})$ rate if $\mathcal{Q}$ nonparametric with metric entropy at most $1/t^2$ [61]. In Section 5, we present an orthogonal estimator that is both robust to slower rates of Q and achieves semiparametric efficiency.

4 Robust Transition Kernels and Estimation with Minimax Learning

Before we present our orthogonal estimator, we must be familiar with another nuisance function: the robust visitation density ratio, *a.k.a.* robust w-function [2, 34]. In this section, we identify the best- and worst-case transition kernels $U^{\pm}$ in our uncertainty set $\mathcal{U}(P)$. Then, we propose a minimax estimator [59] for the robust w-function, which is an important nuisance function for our orthogonal estimator in Section 5.

Identification of $U^{\pm}$. The robust transition kernel $U^{\pm}$ is defined as the feasible perturbed kernel that achieves the sup/inf in the robust Bellman equation $Q^{\pm}(s,a) = \mathcal{T}^{\pm}_{\mathsf{rob}}Q^{\pm}(s,a)$. Let $F^{\pm}(y \mid s,a) = P(V^{\pm}(s') \leq y \mid s,a)$ be the next-state pushforward measure of the robust value function $V^{\pm}$. Then, $U^{\pm}$ is a convex combination of the nominal kernel P and a reweighting of P by an indicator function.

Lemma 4.1. *Suppose $F^{\pm}(\beta^{\pm}_{\tau}(s,a) \mid s,a) = \frac{1}{2} \pm (\frac{1}{2} - \tau)$, where $\beta^{\pm}_{\tau}(s,a)$ is the upper/lower τ-th quantile of $F^{\pm}(\cdot \mid s,a)$. Then,*

$$U^{\pm}(s' \mid s,a)/P(s' \mid s,a) = \Lambda^{-1}(s,a) + (1-\Lambda^{-1})\tau(s,a)^{-1}\mathbb{I}[\pm(V^{\pm}(s') - \beta^{\pm}_{\tau}(s,a)) \geq 0]. \quad (4)$$

The proof idea decompose U into the nominal and perturbed components and use the primal solution of CVaR_{τ} [3]; we formalize this in Section G.2.

Identification of $w^{\pm}$. Using the identification of $U^{\pm}$ in Lemma 4.1, we can now identify the robust w-function based on the Bellman flow equations in the best- or worst-case MDP. In particular, define $d^{\pm,\infty}(s)$ as the γ-averaged visitation of π_{t} under the MDP with transition kernel $U^{\pm}$, *i.e.*, $d^{\pm,\infty}(s) := (1-\gamma)\sum_{h=1}^{\infty}\gamma^{h-1}d^{\pm}_h(s)$, where $d^{\pm}_h$ is the h-th step visitation of π_{t} with the kernel $U^{\pm}$ starting from d_1. Then, the Bellman flow in the robust MDP is $d^{\pm,\infty}(s) = (1-\gamma)d_1(s) + \gamma\mathbb{E}_{\widetilde{s}\sim d^{\pm,\infty},\widetilde{a}\sim\pi_{\mathrm{t}}(\widetilde{s})}U^{\pm}(s \mid \widetilde{s},\widetilde{a})$. Thus, the robust visitation density, defined as $w^{\pm}(s) := {}^{\mathrm{d}d^{\pm,\infty}(s)}/_{\mathrm{d}\nu(s)}$, satisfies the following moment condition for all $f : \mathcal{S} \mapsto \mathbb{R}$:

$$\mathbb{E}[w^{\pm}(s)f(s)] = (1-\gamma)\mathbb{E}_{d_1}[f(s_1)] + \gamma\mathbb{E}[w^{\pm}(s,a)\mathbb{E}_{s'\sim U^{\pm}(s,a)}[f(s')]], \quad (5)$$

where we abused notation and use $w^{\pm}(s,a) := w(s) \cdot {}^{\pi_{\mathrm{t}}(a|s)}/_{\nu(a|s)}$.

4.1 Estimating $w^{\pm}$ with Robust Minimax Indirect Learning

We now propose a penalized minimax estimator for $w^{\pm}$ that generalizes the Minimax Indirect Learning (MIL) of [59] to our robust setting. Our estimator RobustMIL (Algorithm 2) uses a general function class $\mathcal{W} \subset \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}_+$ to approximately solve the moment equation in Eq. (5) by ensuring the difference between the left and right hand side is small for sufficiently many adversaries f in a discriminator class $\mathcal{F} \subset \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$. Since $U^{\pm}$ is unknown, we approximate it via Eq. (4) by plugging in a threshold $\widetilde{\zeta}(s,a,s')$ in the indicator function to approximate the true threshold $\zeta^{\pm}(s,a,s') := V^{\pm}(s') - \beta^{\pm}_{\tau}(s,a)$. This yields the minimax objective in Eq. (6), where we also allow for an optional regularization of the adversary's norm which can be useful for obtaining fast convergence rates.

We make the following standard assumptions for MIL [59].

Assumption 4.2 (Regularity). (i) $\sup_{w \in \mathcal{W} \cup \{w^\pm\}} \|w\|_\infty < \infty$; (ii) the marginal CDF of $V^\pm(s') - \beta^\pm(s,a)$, *i.e.*, $F(y) = P(V^\pm(s') - \beta^\pm_{\tau(s,a)}(s,a) \le y)$, is boundedly differentiable around 0.

We also posit the adversary class is rich enough to capture all projected errors under the adjoint of the operator $\mathcal{J}_{U^\pm} f(s,a) := \gamma \mathbb{E}_{U^\pm}[f(s', \pi_t) \mid s, a] - f(s,a)$.

Assumption 4.3 ($w^\pm$-realizability and completeness). $w^\pm \in \mathcal{W}$ and $\mathcal{J}'_{U^\pm}(\mathcal{W} - w^\pm) \subset \mathcal{F}$.

Finally, we posit standard regularity assumptions on $\mathcal{F}$ for employing the critical radius: we assume it is star-shaped and symmetric (see Appendix D) and satisfies $\sup_{f \in \mathcal{F}} \|f\| / \|\mathcal{T}_{U^\pm} f\| < \infty$.

Theorem 4.4. *Let $\varepsilon_n^{\mathcal{W}}$ denote the maximum critical radii of the following classes:*

$$\mathcal{G}_1 = \{(s,a,s') \mapsto (f(s,a) - \gamma f(s', \pi_t)), f \in \mathcal{F}\},$$
$$\mathcal{G}_2 = \big\{(s,a,s') \mapsto (w(s,a) - w^\pm(s,a))(\gamma f(s', \pi_t) - f(s,a)), f \in \mathcal{F}, w \in \mathcal{W}\big\}.$$

Under Assumptions 4.2 and 4.3, RobustMIL ensures that for any δ, w.p. $1 - \delta$,

$$\big\|\mathcal{J}'_{U^\pm}(\widehat{w} - w^\pm)\big\|_2 \lesssim \varepsilon_n^{\mathcal{W}} + \|\widetilde{\zeta}^\pm - \zeta^\pm\|_\infty + \sqrt{\log(1/\delta)/n}.$$

As before, the critical radius $\varepsilon_n^{\mathcal{W}}$ converges at a $\widetilde{\mathcal{O}}(n^{-1/2})$ rate for parametric classes. Notably, our bounds degrade linearly w.r.t. the ℓ_∞ error in $\widetilde{\zeta}^\pm$ for estimating $\zeta^\pm$. For example, if $\widetilde{\zeta}(s,a,s') = \widehat{v}(s') - \widehat{\beta}(s,a)$ where $\widehat{v}, \widehat{\beta}$ are estimated with RobustFQE, then the ζ-error can be bounded by $\mathcal{O}(\|\widehat{v} - v^\pm\|_\infty + \|\widehat{\beta} - \beta^\pm\|_\infty)$. The proof is in Appendix I, where we prove a more general bound that is gracefully agnostic to the realizability and completeness assumptions.

5 Orthogonal and Efficient Estimator for Robust Policy Value

In this section, we propose an orthogonal estimator that is robust against errors in the nuisances (exhibiting only second-order sensitivity), achieves semiparametric efficiency, and enables inference. Our estimator is based on the efficient influence function (EIF) of $V_{d_1}^\pm$, which is the canonical gradient of a statistical estimand [57]. The adoption of EIFs for developing efficient estimators is a broadly employed technique in causal inference [15, 37] and reinforcement learning [31, 34].

We define the collection of nuisance parameters by $\eta^\pm = (w^\pm, q^\pm, \beta^\pm)$. The notation $\widehat{\eta}$ indicates that these functions are estimated from data, while η or η^* represent their true values.

Theorem 5.1 ((Recentered) Efficient Influence Function). *The (R)EIF of $V_{d_1}^\pm$ is given by:*

$$\psi(s,a,s';\eta^\pm) = V_{d_1}^\pm + w^\pm(s,a)\big(r(s,a) + \gamma \rho^\pm(s,a,s';v^\pm,\beta^\pm) - q^\pm(s,a)\big), \quad \textit{where}$$
$$\rho^\pm(s,a,s';v^\pm,\beta^\pm) = \Lambda(s,a)^{-1} v^\pm(s') + (1 - \Lambda(s,a)^{-1})\big(\beta^\pm(s,a) + \tau^{-1}(v^\pm(s') - \beta^\pm(s,a))_\pm\big).$$

Remark 5.2. When $\Lambda = 1$, there is no shift in the target environment, and the weight on the CVaR term is zero. The (R)EIF then reduces to the (R)EIF in [34] for regular OPE with an infinite horizon. As $\Lambda \to \infty$, the CVaR term becomes predominant, with the quantiles $\beta^\pm(s,a)$ taking extreme values. This yields the (novel) (R)EIF for the problem in [22], where the expected value term is replaced solely by a CVaR component in the Bellman equation.

The (R)EIF forms the basis of our orthogonal estimator. First, we note that $\mathbb{E}[\psi(s,a,s';\eta^\pm)]$ is an unbiased estimator of $V_{d_1}^\pm$. Furthermore, the expression for $\psi(s,a,s';\eta^\pm)$ depends only on quantities $w^\pm, q^\pm, \beta^\pm$ which can be estimated from data. Thus, we can cast the expression $\mathbb{E}[\psi(s,a,s';\eta^\pm)]$ as a statistical estimand to be learned from the observed distribution. This suggests a natural two-stage estimator that we summarize in Algorithm 3. In the first stage, we estimate the nuisance parameters $\widehat{\eta}$ from the data with K-fold cross-fitting; in the second stage, these estimates are incorporated into the (R)EIF expression and we calculate the empirical average using the observed data. We summarize our procedure in Algorithm 3.

Algorithm 3 Orthogonal Estimator for $V_{d_1}^{\pm}$

```
Input: Dataset $\mathcal{D}$, number of splits $K$.
for $k = 1, 2, \dots, K$ do
    Use data $\mathcal{D} \setminus \mathcal{D}_k$ to learn $(q^{\pm,[k]}, \beta^{\pm,[k]})$ with Algorithm 1 and $w^{\pm,[k]}$ with Algorithm 2
    for $i = \lfloor (k-1)n/K \rfloor, \dots, \lfloor kn/K \rfloor - 1$ do $\psi_i^{\pm} = \psi(s_i, a_i, s_i', \widehat{\eta}^{\pm})$
Output: $\widehat{V}_{d_1}^{\pm} = \frac{1}{n} \sum_{i=1}^{n} \psi_i^{\pm}$.
```

The nuisance estimation is detailed in Sections 3.2 and 4.1. The reliance on the EIF confers our estimator desirable statistical properties including a second order bias due to the nuisances, meaning the bias has a product structure with respect to the nuisance errors. Thus, this special structure orthogonalizes away the dependency on $\widehat{Q}^{\pm}$ errors which now only appear in second order. Furthermore, our estimator is semiparametrically efficient in the sense that under mild consistency assumptions, it achieves minimum variance among all regular and asymptotically linear (RAL) estimators. We provide theoretical justifications for these properties in the next section.

5.1 Theoretical Guarantees of the Orthogonal Estimator

We now characterize the theoretical properties of our orthogonal estimator. We consider the K-fold cross-fitted estimator in Algorithm 3 given by

$$\widehat{V}_{d_1}^{\pm} = \tfrac{1}{n} \textstyle\sum_{k=1}^{K} \sum_{(s,a,s') \in \mathcal{D}^k} \psi(s, a, s'; \widehat{\eta}^{[k]}),$$

where nuisances $\widehat{\eta}^{[k]}, k \in [K]$ are trained on all data excluding the k^{th} fold $\mathcal{D}^k$. The following theorem outlines the theoretical guarantees of this estimator:

Theorem 5.3 (Efficiency of $\widehat{V}^{\pm}$). *Let $r_{n,p}^{w}, r_{n,p}^{q}, r_{n,p}^{\beta}$ be functions of $n = |\mathcal{D}|$ such that $\|\mathcal{J}_{U^{\pm}}'(\widehat{w}^{\pm,[k]} - w^*)\|_p \le r_{n,p}^{w}$, $\|\widehat{q}^{\pm,[k]} - q^*\|_p \le r_{n,p}^{q}$, and $\|\beta^{\pm,[k]} - \beta^*\|_p \le r_{n,p}^{\beta}$ for any $k \in [K]$. Furthermore, assume that the regularity conditions in Assumption 4.2 hold. Then:*

$$|\widehat{V}_{d_1} - V_{d_1}^*| \lesssim O_p(n^{-1/2}) + O_p(r_{n,2}^{w} r_{n,2}^{q} + (r_{n,\infty}^{q})^2 + (r_{n,\infty}^{\beta})^2) \qquad \text{(Rates)}$$

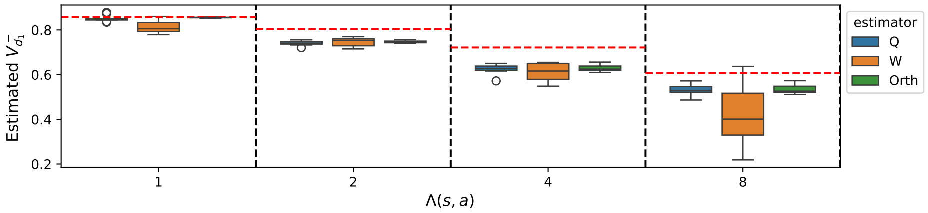
Furthermore, if $r_{n,2}^{w} \vee r_{n,2}^{q} = o_p(1)$, $r_{n,2}^{w} r_{n,2}^{q} = o_p(n^{-1/2})$, $r_{n,\infty}^{q} = o_p(n^{-1/4})$, and $r_{n,\infty}^{\beta} = o_p(n^{-1/4})$, then $\widehat{V}_{d_1}$ satisfies:

$$\sqrt{n}(\widehat{V}_{d_1} - V_{d_1}^*) \xrightarrow{d} \mathcal{N}(0, \Sigma), \quad \Sigma = \mathrm{Var}(\psi(s, a, s'; \eta)). \qquad \text{(Normality \& Efficiency)}$$

Moreover, Σ is the minimum achievable asymptotic variance among RAL estimators in the nonparametric model for (s, a, s') (the efficiency bound).

We provide the intuition along with a detailed proof in Appendix E. The first part of Theorem 5.3 implies that as long as we estimate the nuisances at rates faster that $n^{-1/4}$, then we can learn $\widehat{V}_{d_1}^{\pm}$ at parametric rates. The second part of Theorem 5.3 states that under mild consistency assumptions, our estimator attains the efficiency bound and is asymptotically normal. That means, for example, we can construct asymptotically valid lower 95%-confidence bound on $\widehat{V}_{d_1}^{-}$ by simply subtracting 1.64 times $\widehat{\mathrm{se}} = \frac{1}{n}\left(\sum_{k=1}^{K} \sum_{(s,a,s') \in \mathcal{D}^k} (\psi(s, a, s'; \widehat{\eta}^{[k]}) - \widehat{V}_{d_1}^{-})^2\right)^{1/2}$. Then, we can be sure to have a bound on the worst-case RL policy value, accounting *both* for potential environment shift and finite data. Finally, in Appendix J, we describe two settings when our orthogonal estimator remains valid even if some nuisances are *inconsistent*, which is a desirable guarantee for sensitivity analysis [20].

Putting everything together. We can instantiate Theorem 5.3 with the nuisance estimators from the previous sections. First, use RobustFQE to estimate $\widehat{q}^{\pm}$ and $\widehat{\beta}^{\pm}$, ensuring $\|\widehat{q}^{\pm} - Q^{\pm}\|_2 \le \mathcal{O}(\varepsilon_n^{\mathcal{Q}} + \mathrm{err}_{\mathsf{QR}}^2)$. Under smoothness conditions (Lemma C.2), the L_2 guarantee for $\widehat{q}^{\pm}$ implies an L_∞ guarantee for $\widehat{q}^{\pm}$, which also ensures an L_∞ guarantee for $\widehat{\beta}^{\pm}$. This ensures $\max(\|\widehat{q}^{\pm} - Q^{\pm}\|_\infty, \|\widehat{\beta}^{\pm} - \beta^{\pm}\|_\infty)$ is well-controlled. Then, we can set $\widetilde{\zeta}^{\pm}(s, a, s') = \widehat{q}^{\pm}(s', \pi_{\mathsf{t}}) - \widehat{\beta}^{\pm}(s, a)$ and run RobustMIL for estimating $\widehat{w}^{\pm}$. By Theorem 4.4, its projected-L_2 error is $\mathcal{O}(\varepsilon_n^{\mathcal{W}} + \|\widehat{q}^{\pm} - Q^{\pm}\|_\infty + \|\widehat{\beta}^{\pm} - \beta^{\pm}\|_\infty)$. Therefore, the final rate via Theorem 5.3 is $\mathcal{O}((\varepsilon_n^{\mathcal{Q}} + \mathrm{err}_{\mathsf{QR}}^2) \cdot \varepsilon_n^{\mathcal{W}} + \|\widehat{q}^{\pm} - Q^{\pm}\|_\infty^2 + \|\widehat{\beta}^{\pm} - \beta^{\pm}\|_\infty^2)$.



Λ	Mean squared error (MSE) to true worst-case policy value		
	Q	**W**	**Orth**
1	$.000240 \pm .000170$	$.002722 \pm .002266$	$\mathbf{.000005 \pm .000006}$
2	$.004053 \pm .001329$	$.003584 \pm .002311$	$\mathbf{.003244 \pm .000600}$
4	$.009799 \pm .005172$	$.013862 \pm .009228$	$\mathbf{.008721 \pm .002543}$
8	$.006247 \pm .003980$	$.052643 \pm .050839$	$\mathbf{.005713 \pm .002730}$

Figure 1: Results of our synthetic data experiments. We show results for our three estimators on all four Λ values, over our 10 experiment replications. **Above:** Box plot summarizing range of policy value estimates for each combination of estimator and Λ, with Horizontal red dashed lines showing the true worst-case policy values $V^-_{d_1}$. **Below:** Table summarizing the corresponding MSE of these estimators for the true worst-case policy value, along with one standard deviation errors.

6 Empirical Evaluation

We now provide a proof-of-concept empirical investigation of our theory. We experiment with our proposed methodology in a simple synthetic environment. We first discuss our environment, then we discuss our approach for solving for the nuisances functions $\eta^\pm$, and finally we provide some empirical results of our corresponding Orthogonal estimator, and compare these to corresponding weighted or direct estimators using the $Q^\pm$ or $w^\pm$ nuisances only.

Experimental Setup We consider a synthetic MDP with a one-dimensional state and two actions, which is based on a simple control problem with non-deterministic dynamics. We consider the task of estimating the worst-case policy value $V^-_{d_1}$ of a fixed candidate policy π_t, for four different fixed sensitivity parameter values: $\Lambda(s,a) \in \{1,2,4,8\}$.

We considered estimation using the following methods:

1. **Q** (RobustFQE): Direct method using the estimated robust quality function $\widehat{Q}^-$ only.
2. **W** (RobustMIL): Importance-sampling method using the estimated robust density ratio $\widehat{w}^-$ only.
3. **Orth**: Our orthogonal estimator which combines the former two, as described in Algorithm 3.

We performed 10 replications of our experimental procedure, in each case: (1) sampling a dataset 20,000 tuples using a different fixed logging policy π_b; (2) fitting the nuisance functions Q^-, β^-, and w^- along the lines of Algorithms 1 and 2 for each Λ; and (3) for each Λ estimating the corresponding robust policy value $V^-_{d_1}$ for all estimators using our estimated nuisances.

Full experimental details, including our MDP, target/logging policies, methodology for computing the true robust policy values $V^-_{d_1}$, and nuisance estimation, are provided in Appendix K.

Results We summarize our results in Fig. 1. We note that all of our estimators are almost always valid for all values of Λ that we experimented with. **Orth** consistently has the lowest mean squared error for the true worst-case policy value. In particular, incorporating the robust importance-sampling weights improves the RobustFQE estimator **Q**, even though these importance-sampling weights by themselves (as in **W**) are much noiser estimators. This is consistent with our theory that the orthogonal estimator is semiparametrically efficient and insensitive to errors in the nuisance functions.

7 Conclusion

We consider the problem of infinite-horizon OPE in RL settings when there can be unknown, but bounded, shifts in the transition distribution compared to the transition distribution generating the data. This can, for example, occur when there is unobserved confounding so we the transitions we see do not reflect true causal ones, there is non-stationarity in the environment we encounter, or we encounter adversarial environments. We consider a sensitivity model for such transition kernel shifts analogous to the classic MSM for static decision making, and provided theory for identifying and estimating the sharp (*i.e.*, tightest possible) bounds on the best/worst-case policy value, as well as the corresponding robust Q- and state density ratio functions. Our estimator for best/worst-case policy value is orthogonal (it is insensitive to how we estimate the nuisance functions) and achieves semiparametric efficiency (it achieves the best possible asymptotic variance). It also enables inference so we can be sure to get a bound for the robust policy value given finite data.

References

[1] Alekh Agarwal, Yuda Song, Wen Sun, Kaiwen Wang, Mengdi Wang, and Xuezhou Zhang. Provable benefits of representational transfer in reinforcement learning. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2114–2187. PMLR, 2023.

[2] Philip Amortila, Dylan J Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning. *arXiv preprint arXiv:2401.09681*, 2024.

[3] Marcus Ang, Jie Sun, and Qiang Yao. On the dual representation of coherent risk measures. *Annals of Operations Research*, 262:29–46, 2018.

[4] Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers under the margin condition. *arXiv preprint math/0507180*, 2005.

[5] Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pages 511–520. PMLR, 2021.

[6] Alexandre Belloni, Victor Chernozhukov, Ivan Fernandez-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

[7] Andrew Bennett, Nathan Kallus, and Miruna Oprescu. Low-rank mdps with continuous action spaces. *arXiv preprint arXiv:2311.03564*, 2023.

[8] Pallab K Bhattacharya and Ashis K Gangopadhyay. Kernel and nearest-neighbor estimation of a conditional quantile. *The Annals of Statistics*, pages 1400–1415, 1990.

[9] Matteo Bonvini, Edward Kennedy, Valerie Ventura, and Larry Wasserman. Sensitivity analysis for marginal structural models. *arXiv preprint arXiv:2210.04681*, 2022.

[10] Haïm Brezis and Petru Mironescu. Where sobolev interacts with gagliardo–nirenberg. *Journal of functional analysis*, 277(8):2839–2864, 2019.

[11] David Bruns-Smith and Angela Zhou. Robust fitted-q-evaluation and iteration under sequentially exogenous unobserved confounders. *arXiv preprint arXiv:2302.00662*, 2023.

[12] David 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.

[13] Domagoj Ćevid, Loris Michel, Jeffrey Näf, Nicolai Meinshausen, and Peter Bühlmann. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *arXiv preprint arXiv:2005.14458*, 2020.

[14] Jonathan Chang, Kaiwen Wang, Nathan Kallus, and Wen Sun. Learning bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning*, pages 2938–2971. PMLR, 2022.

[15] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

[16] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.

[17] Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, 2022.

[18] Victor Chernozhukov, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. Automatic debiased machine learning for dynamic treatment effects and general nested functionals. *arXiv preprint arXiv:2203.13887*, 2022.

[19] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28, 2015.

[20] Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657, 2023.

[21] Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*, 2021.

[22] Yihan Du, Siwei Wang, and Longbo Huang. Provably efficient risk-sensitive reinforcement learning: Iterated cvar and worst path. In *The Eleventh International Conference on Learning Representations*, 2022.

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

[24] Anouar El Ghouch and Marc G Genton. Local polynomial quantile regression with parametric features. *Journal of the American Statistical Association*, 104(488):1416–1429, 2009.

[25] Kevin Elie-Dit-Cosaque and Véronique Maume-Deschamps. Random forest estimation of conditional distribution functions and conditional quantiles. *Electronic Journal of Statistics*, 16 (2):6553–6583, 2022.

[26] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

[27] Vineet Goyal and Julien Grand-Clement. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023.

[28] Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3):292–304, 2022.

[29] Hidehiko Ichimura and Whitney K Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022.

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

[31] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

[32] Nathan Kallus. What's the harm? sharp bounds on the fraction negatively affected by treatment. *Advances in Neural Information Processing Systems*, 35:15996–16009, 2022.

[33] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *The Journal of Machine Learning Research*, 21(1): 6742–6804, 2020.

[34] Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research*, 70(6):3282–3302, 2022.

[35] Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in neural information processing systems*, 33:22293–22304, 2020.

[36] Edward H Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

[37] Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

[38] J Kolter. The fixed points of off-policy td. *Advances in Neural Information Processing Systems*, 24, 2011.

[39] Navdeep Kumar, Kfir Levy, Kaixin Wang, and Shie Mannor. Efficient policy iteration for robust markov decision processes via regularization. *arXiv preprint arXiv:2205.14327*, 2022.

[40] Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Yehuda Levy, and Shie Mannor. Policy gradient for rectangular robust markov decision processes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=NLpXRrjpa6.

[41] Mark J Laan and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.

[42] Giovanni Leoni. *A first course in Sobolev spaces*. American Mathematical Soc., 2017.

[43] Greg Lewis and Vasilis Syrgkanis. Double/debiased machine learning for dynamic treatment effects. In *NeurIPS*, pages 22695–22707, 2021.

[44] Shie Mannor, Ofir Mebel, and Huan Xu. Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

[45] Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

[46] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[47] Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

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

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

[50] Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35:32211–32224, 2022.

[51] Jeffrey S Racine and Kevin Li. Nonparametric conditional quantile estimation: A locally weighted quantile kernel approach. *Journal of Econometrics*, 201(1):72–94, 2017.

[52] R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

[53] Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

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

[55] Ichiro Takeuchi, Quoc Le, Timothy Sears, Alexander Smola, et al. Nonparametric quantile estimation. 2006.

[56] Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

[57] Anastasios A Tsiatis. Semiparametric theory and missing data. 2006.

[58] John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

[59] Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

[60] Mark J van der Laan, Sherri Rose, Wenjing Zheng, and Mark J van der Laan. Cross-validated targeted minimum-loss-based estimation. *Targeted learning: causal inference for observational and experimental data*, pages 459–474, 2011.

[61] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

[63] Jie Wang, Rui Gao, and Hongyuan Zha. Reliable off-policy evaluation for reinforcement learning. *Operations Research*, 72(2):699–716, 2024.

[64] Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

[65] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[66] Wenhao Xu, Xuefeng Gao, and Xuedong He. Regret bounds for Markov decision processes with recursive optimized certainty equivalents. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38400–38427. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/xu23d.html.

Appendices

A Notations

Table 1: List of Notations

$\mathcal{S}, \mathcal{A}$	State and action spaces.
$\Delta(S)$	The set of distributions supported by set S.
d_1	The initial state distribution.
$\Lambda(s,a)$	Tolerance parameter for kernel shift at (s,a). Takes values $[1,\infty]$.
$\tau(s,a)$	$\tau(s,a) = \frac{1}{1+\Lambda(s,a)} \in [0, \frac{1}{2}]$.
$V^{\pm}, Q^{\pm}$	Robust value and quality functions of the target policy π_{t}.
$f(s,\pi)$	$f(s,\pi) := \mathbb{E}_{a\sim\pi(s)}[f(s,a)]$.
$U^{\pm}(s' \mid s,a)$	Robust transition kernel which attains the best- or worst-case value.
$\mathcal{T}_U, \mathcal{T}^{\pm}_{\mathsf{rob}}$	Bellman operator under U and the robust Bellman operators.
$\mathcal{J}_U$	$\mathcal{J}_U f(s,a) := \gamma \mathbb{E}_U[f(s',\pi_{\mathsf{t}}) \mid s,a] - f(s,a)$
$\beta^{\pm}_{\tau}(s,a)$	The upper τ-th quantile of $V^{+}(s')$ and lower τ-th quantile of $V^{-}(s')$, $s' \sim P(s,a)$.
$d^{\pi_{\mathsf{t}},\infty}_{d_1,U}$	The γ-discounted average visitation of π_{t} under MDP with transition U starting from d_1.
$d^{\pm,\infty}$	$d^{\pm,\infty} = d^{\pi_{\mathsf{t}},\infty}_{d_1,U^{\pm}}$.
$\nu(s), \nu(s,a)$	Data generating distribution. $\nu(s)$ marginalizes over actions.
$w^{\pm}$	$w^{\pm} = {}^{\mathrm{d}d^{\pm,\infty}}/_{\mathrm{d}\nu}$. This is valid both as a function of s or (s,a).
$\omega(s,a)$	$\omega(s,a) = \frac{\pi_{\mathsf{t}}(a\|s)}{\nu(a\|s)}$.
x_+, x_-	$\max(0,x), \min(0,x)$ respectively, for $x \in \mathbb{R}$.
$x \lesssim y$	$x \leq Cy$ for some constant C.
$\mathbb{E}_n$	Empirical average over n samples.
$\\|f\\|_p$	L^p norm, $(\mathbb{E}\|f(X)\|^p)^{1/p}$.
$f^{\star}$	True (oracle) value of a parameter or function f.
$f, \bar{f}$	Putative value of a parameter or function f.
$\widehat{f}$	Estimated value of a parameter or function f.

B Other Related Works

Other Robust MDPs. There is a rich literature on Robust MDPs [27, 30, 44, 65] with s, a-rectangular uncertainty sets, but these foundational works assumed knowledge of the transition kernel. Recently, learning-based robust MDP algorithms have been proposed for uncertainty sets under the total variation [40, 50] and more generally L_p balls [39]. These L_p uncertainty sets are additive in nature, *i.e.*, the adversary adds or subtracts a vector in the ℓ_p ball to $P(\cdot \mid s, a)$, whereas our uncertainty set is multiplicative in nature, *i.e.*, the adversary can multiply or divide a bounded factor and is more commonly used in causal inference to model unobserved confounding.

Risk-sensitive RL with Dynamic Risk. RL with dynamic risk measures (*a.k.a.* iterated risk measures) is tightly related and often equivalent to Robust MDPs [19]. In Lemma 3.1, we show that our MSM uncertainty set is indeed equivalent to iterated risk RL with the risk measure $\Lambda\mathbb{E} + (1-\Lambda)\,\mathrm{CVaR}_\tau$. Efficient online RL algorithms have been proposed for similar measures [22, 66]. However, our focus is on deriving the optimal OPE estimators for the problem, which involves a different set of challenges such as deriving the efficiency bound and ensuring sharpness guarantees even when nuisances are estimated slowly.

C Higher Order Norms via Smoothness

For any $x \in \mathbb{R}^+$, define $\lfloor x \rfloor$ as the greatest integer that is strictly less than x, and let x and $\{x\} = x - \lfloor x \rfloor$ represent the fractional part. Thus, we obtain the distinct decomposition $x = \lfloor x \rfloor + \{x\}$, where $\lfloor x \rfloor \in \mathbb{N}$ and $\{x\} \in (0, 1]$.

Definition C.1 (α-smooth functions)**.** Given $\alpha \in (0, \infty)$ and $\mathcal{X} \subseteq \mathbb{R}^m$, $f : \mathcal{X} \to \mathbb{R}$ is an α-smooth function if (1) the mixed derivatives up to $\lfloor \alpha \rfloor$-order exist and are bounded; and (2) all $\lfloor \alpha \rfloor$-order derivatives are $\{\alpha\}$-Hölder continuous [42].

Lemma C.2 (L^∞ Bound for α-Smooth Functions)**.** *Let $f : \mathcal{X} \to \mathbb{R}, \mathcal{X} \subseteq \mathbb{R}^m$ be an α-smooth function as in Definition C.1. Then, if $\mathcal{X}$ is $\mathbb{R}^m$, a half-space or a bounded Lipschitz domain in $\mathbb{R}^m$, there exists a constant C such the following inequality holds:*

$$\|f\|_\infty \leq C \|f\|_p^{\frac{p\alpha}{p\alpha + m}}.$$

Proof. This lemma is a direct application of the fractional Gagliardo-Nirenberg interpolation inequality (Theorem 1 in [10]) from the functional analysis literature. For a more comprehensive exposition on this result, see Appendix A.1 in [7]. □

D Localized Rademacher Complexity and Critical Radius

In this section, we recap the localized Rademacher complexity and critical radius, which is a standard complexity measure for obtaining fast rates for squared loss [62]. Let $\mathcal{G}$ be a class of functions $g : \mathcal{Z} \to \mathbb{R}$. Given n datapoints $z_1, z_2, \dots, z_n$, the empirical localized Rademacher complexity is:

$$\mathcal{R}_n(\varepsilon, \mathcal{G}) := \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G} : \|g\|_n \leq \varepsilon} \frac{1}{n} \sum_{i=1}^n \epsilon_i g(z_i) \right],$$

where $\mathbb{E}_\sigma$ is expectation over n independent Rademacher random variables $\sigma_1, \sigma_2, \dots, \sigma_n$, *i.e.*, $\mathbb{E}_\sigma[\cdot] = \frac{1}{2^n} \sum_{\sigma \in \{-1,1\}^n} [\cdot]$. Note that when $\varepsilon = \infty$, there is no localization and $\mathcal{R}_n(\infty, \mathcal{G})$ reduces to the vanilla Rademacher complexity. Let $C := \sup_{g \in \mathcal{G}} \|g\|_\infty$ be the envelope of $\mathcal{G}$. Then, the critical radius of $\mathcal{G}$ with n, called ε_n, is the smallest ε that satisfies $\mathcal{R}_n(\varepsilon, \mathcal{G}) \leq \varepsilon^2/C$.

Unless otherwise stated, we will posit that $\mathcal{G}$ is star-shaped: there exists $g_0 \in \mathcal{G}$ such that for all $g \in \mathcal{G}$ and $\alpha \in [0, 1]$, we have $\alpha g_0 + (1-\alpha) g \in \mathcal{G}$. If not, we can replace $\mathcal{G}$ by its star-hull, *i.e.*, the smallest star-shaped set containing $\mathcal{G}$. We will also posit that $\mathcal{G}$ is symmetric for simplicity.

The critical radius is a well-studied quantity in statistics [62] and also recently in RL [23, 59]. For example if $\mathcal{G}$ has d VC-subgraph dimension, then w.p. $1-\delta$, $\varepsilon_n \leq \mathcal{O}(\sqrt{d\log n/n})$. For nonparametric models with metric entropy at most $1/t^{\beta}$, the critical radius can also be bounded by $\mathcal{O}(n^{-1/(\max(2+\beta,2\beta))})$ [59], *e.g.*, is $\mathcal{O}(n^{-1/4})$ if $\beta = 2$.

E Proof of Theorem 5.3

E.1 Intuition for Theorem 5.3

We provide some intuition for the results in Theorem 5.3. Consider the V^- bound and let us decouple the indicator $\mathbb{I}\left[v(s') - \beta(s,a) \leq 0\right]$ that appears implicitly in the $(v^-(s') - \beta^-(s,a))_-$ notation of Theorem 5.1. We augment the set of nuisances with $\zeta(s,a,s') = v^-(s') - \beta^-(s,a)$ such that $(v^-(s') - \beta^-(s,a))_- = (v^-(s') - \beta^-(s,a))\mathbb{I}\left[\zeta(s,a,s') \leq 0\right]$. We state the following lemma (which we elaborate upon in Lemmas E.4 and E.5 in the Appendix):

Lemma E.1 (Double sharpness with correct $\zeta^\star$)**.** *Let* $\mathbb{E}[\psi(s,a,s';q,w,\beta,\zeta^\star)]$ *be the expectation of the (R)EIF with an arbitrary nuisance set* $\eta = (w,q,\beta)$*, but where the indicator* $\mathbb{I}[v^-(s') \leq \beta^-(s,a)]$ *has been replaced with the correct indicator* $\mathbb{I}[\zeta^\star(s,a,s') \leq 0]$*. Then:*

$$V^-_{d_1} = \mathbb{E}[\psi(s,a,s';q,w^\star,\beta^\star,\zeta^\star)] = \mathbb{E}[\psi(s,a,s';q^\star,w,\beta^\star,\zeta^\star)]$$

This lemma implies that if $\beta^- = (\beta^*)^-$ and $\zeta = \zeta^*$, then the estimator $\widehat{V}^-_{d_1}$ has a property known as “double-robustness" [37] or “double-sharpness" [21] in q and w, meaning the bias vanishes when either q or w is consistent. Moreover, the convergence rate would be $O_p(r^w_{n,2}r^q_{n,2})$. This condition holds provided that β and ζ are correctly specified. However, estimation errors in β introduce an additional $O_p\left((r^{\beta}_{n,\infty})^2\right)$ term, reflecting that β is first-order optimal for the CVaR component. Additionally, discrepancies between ζ and ζ^* contribute an extra $O_p\left((r^{q}_{n,\infty})^2\right)$ to the error. While this discussion gives some insight into how we achieve the results in Theorem 5.3, we provide a a rigorous analysis in the next section.

E.2 The Proof

For this proof, our focus will be on $\widehat{V}^-_{d_1}$. The argument for $\widehat{V}^+_{d_1}$ is analogous, following a symmetric approach. To improve the clarity of our exposition, we will omit the $-$ and τ indices, assuming their presence is clear from the context.

For simplicity, we assume that n is a multiple of K such that $n = Kn_K$, where n_K is the size of a fold. We let $\mathbb{E}_n, \mathbb{E}_k$ denote the empirical averages over the entire sample and the k^{th} fold, respectively. Recall that we use $\widehat{\eta} = (\widehat{w}, \widehat{q}, \widehat{\beta})$ and $\eta^* = (w^*, q^*, \beta^*)$ to denote the estimated and oracle nuisances, respectively.

We further suppress the dependency on s, a in Λ and τ and we write the ρ term in Theorem 5.1 as

$$\rho(s,a,s';v,\beta) = (1-\lambda)v(s') + \lambda\big(\beta(s,a) + \tau^{-1}(v(s') - \beta(s,a))_-\big). \tag{7}$$

We justify this by noting that the analysis holds regardless of whether λ and τ depend on s, a. Sometimes, it will be useful to decouple the indicator $\mathbb{I}\left[v(s') - \beta(s,a) \leq 0\right]$ implicit in the definition of ρ. In this case, we augment the set of nuisances with $\zeta(s,a,s') = v(s') - \beta(s,a)$ and write ρ as

$$\rho(s,a,s';v,\beta,\zeta) = (1-\lambda)v(s') + \lambda\big(\beta(s,a) + \tau^{-1}(v(s') - \beta(s,a))\mathbb{I}\left[\zeta(s,a,s') \leq 0\right]\big). \tag{8}$$

Similarly define $\psi(\cdot;w,q,\beta,\zeta)$ with the $\rho(\cdot;v,\beta,\zeta)$.

E.3 Auxiliary Lemmas

Definition E.2 (Margin Condition)**.** A function $f : \mathcal{X} \to \mathbb{R}$ of some random variable X is said to satisfy the margin condition with sharpness $\alpha \in [0,\infty]$ (or more succinctly, an α-margin) if there exist a fixed constant $c > 0$ such that

$$\forall t > 0 : P(0 < |f(X)| \leq t) \leq ct^{\alpha}.$$

If $f(X)$ is either zero or bounded away from zero almost surely, then f satisfies an infinite margin, *i.e.*, $\alpha = \infty$ [32, Lemma 2]. If $f(X)$ is continuously distributed in a neighborhood around 0, *i.e.*, its CDF is boundedly differentiable on $(-\varepsilon, 0) \cup (0, \varepsilon)$ for some $\varepsilon > 0$, then f has a 1-margin [32, Lemma 3].

Lemma E.3 (Margin Guarantees)**.** *For any $f : \mathcal{X} \to \mathbb{R}$ satisfying α-margin (Definition E.2), $p \in [1, \infty]$, and any $g : \mathcal{X} \to \mathbb{R}$, the following statements hold for some constant $C > 0$:*

$$\mathbb{E}[(\mathbb{I}[g(X) \leq 0] - \mathbb{I}[f(X) \leq 0])f(X)] \leq C\|f - g\|_p^{\frac{p(1+\alpha)}{p+\alpha}}, \tag{9}$$

$$P[\mathbb{I}[g(X) \leq 0] \neq \mathbb{I}[f(X) \leq 0], f(X) \neq 0] \leq C\|f - g\|_p^{\frac{p\alpha}{p+\alpha}}, \tag{10}$$

where $\|\cdot\|_p$ is the L^p norm and we set $\infty t/\infty = t$ in the exponents.

The proof of Eq. (9) for any $p \in [1, \infty]$ and of Eq. (10) for $p = \infty$ is given in [4, Lemmas 5.1 and 5.2]. The proof of Eq. (10) for $p < \infty$ is given in [32, Lemma 5].

Lemma E.4 (Sharpness with correct $q^\star$ and $\beta^\star$)**.** $\frac{1}{n}\sum_{(s,a,s')\sim\mathcal{D}} \psi(s, a, s'; w, q, \beta)$ *is an unbiased estimator of $V^\star_{d_1}$ when $q = q^\star, \beta = \beta^\star$, i.e.,*

$$(1 - \gamma)\mathbb{E}_{d_1} v^\star(s_1) = \mathbb{E}[\psi(s, a, s'; w, q^\star, \beta\star)].$$

Proof. Since $q^\star$ and $\beta^\star$ are correct, the robust Bellman equation holds, and so for every s, a,

$$\mathbb{E}\big[(1 - \lambda)v^\star(s') + \lambda(\beta^\star(s, a) + \tau^{-1}(v^\star(s') - \beta^\star(s, a))_-) \mid s, a\big] = 0.$$

Thus, multiplying by any w does not change the fact that the debiasing term in ψ has expectation zero. Since we have $v^\star$, the first term in ψ is exactly the estimand, which concludes the proof. □

Lemma E.5 (Sharpness with correct w^* and ζ^*)**.** $\frac{1}{n}\sum_{(s,a,s')\sim\mathcal{D}} \psi(s, a, s'; w, q, \beta, \zeta)$ *is an unbiased estimator of $V^\star_{d_1}$ when $w = w^\star, \zeta = \zeta^\star$, i.e.,*

$$(1 - \gamma)\mathbb{E}_{d_1} v^\star(s_1) = \mathbb{E}[\psi(s, a, s'; q, w^\star, \beta, \zeta^\star)]$$

Proof. Let $P^\star$ denote the robust transition kernel and let $d^\star$ denote the robust visitation measure under π, which satisfies: for all functions f,

$$\mathbb{E}_{d^\star}[f(s, a)] = (1 - \gamma)\mathbb{E}_{d_1} f(s, \pi) + \gamma\mathbb{E}_{\widetilde{s}, \widetilde{a}\sim d^\star, s\sim P^\star(s,a)}[f(s, \pi)].$$

Since $\zeta^\star$ is correct, for any v, s, a, we have

$$\begin{aligned}
&\mathbb{E}_{s'\sim P(s,a)}\left[(1 - \lambda)v(s') + \lambda\left(\beta(s, a) + \tau^{-1}(v(s') - \beta(s, a))\mathbb{I}\left[\zeta^\star(s, a, s') \leq 0\right]\right)\right] \\
&= \mathbb{E}_{s'\sim P(s,a)}\left[(1 - \lambda)v(s') + \lambda\tau^{-1}v(s')\mathbb{I}\left[\zeta^\star(s, a, s') \leq 0\right]\right] && (\bigstar) \\
&= \mathbb{E}_{s'\sim P^\star(s,a)}[v(s')], && \text{(Lemma 4.1)}
\end{aligned}$$

where in $\bigstar$ we used $\mathbb{E}_{s'\sim P(s,a)}\left[\beta(s, a)\left(1 - \tau^{-1}\mathbb{I}\left[\zeta^\star(s, a, s') \leq 0\right]\right)\right] = \beta(s, a)\left(1 - \tau^{-1}\tau\right) = 0$. That is, for all function f, we have

$$\begin{aligned}
&(1 - \gamma)\mathbb{E}_{d_1} v(s_1) + \mathbb{E}[w^\star(s, a)(r(s, a) + \gamma\rho(s, a, s'; v, \beta, \zeta^\star) - q(s, a))] \\
&= (1 - \gamma)\mathbb{E}_{d_1} v(s_1) + \mathbb{E}_{s,a\sim d^\star}[r(s, a) + \gamma\rho(s, a, s'; v, \beta, \zeta^\star) - q(s, a)] \\
&= \mathbb{E}_{s,a\sim d^\star}[r(s, a)] + (1 - \gamma)\mathbb{E}_{d_1} v(s_1) + \mathbb{E}_{s,a\sim d^\star}\left[\gamma\mathbb{E}_{s'\sim P^\star(s,a)}[v(s')] - q(s, a)\right] \\
&= \mathbb{E}_{s,a\sim d^\star}[r(s, a)] && \text{(robust Bellman flow)} \\
&= (1 - \gamma)\mathbb{E}_{d_1} v^\star(s_1).
\end{aligned}$$

This concludes the proof. □

E.4 Proof of Rates

The estimation error is given by:

$$|\widehat{V}_{d_1} - V^*_{d_1}| = \left|\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_k[\psi(s,a,s';\widehat{\eta}^{[k]})] - V^*_{d_1}\right| \le \frac{1}{K}\sum_{k=1}^{K} \left|\mathbb{E}_k[\psi(s,a,s';\widehat{\eta}^{[k]})] - V^*_{d_1}\right|$$

We wish need to bound $\left|\mathbb{E}_k[\psi(s,a,s';\widehat{\eta}^{[k]})] - V^*_{d_1}\right|$. We have that:

$$\left|\mathbb{E}_k[\psi(s,a,s';\widehat{\eta}^{[k]})] - V^*_{d_1}\right| \le \left|\mathbb{E}_k[\psi(s,a,s';\widehat{\eta}^{[k]})] - \mathbb{E}[\psi(s,a,s';\widehat{\eta}^{[k]})]\right| + \left|\mathbb{E}[\psi(s,a,s';\widehat{\eta}^{[k]})] - V^*_{d_1}\right|$$

The first term is $O_p(n^{-1/2})$ by the CLT. We are now interested in bounding the second term:

$$\varepsilon(\widehat{\eta}) := \left|\mathbb{E}[\psi(s,a,s';\widehat{\eta})] - V^*_{d_1}\right|. \tag{11}$$

where we dropped the $[k]$ indicator without loss of generality. We further decompose $\varepsilon(\widehat{\eta})$ into two error terms, ε_A and ε_B, as follows:

$$\begin{aligned}
\varepsilon(\widehat{\eta}) &= \left|\mathbb{E}\Big[\psi(s,a,s';\widehat{q},\widehat{w},\widehat{\beta})\Big] - \mathbb{E}\Big[\psi(s,a,s';\widehat{q},w^\star,\widehat{\beta},\zeta^\star)\Big]\right| && \text{(Lemma E.5)}\\
&\le \left|\mathbb{E}\Big[\psi(s,a,s';\widehat{q},\widehat{w},\widehat{\beta})\Big] - \mathbb{E}\Big[\psi(s,a,s';\widehat{q},\widehat{w},\widehat{\beta},\zeta^\star)\Big]\right| && (\varepsilon^A)\\
&\quad + \left|\mathbb{E}\Big[\psi(s,a,s';\widehat{q},\widehat{w},\widehat{\beta},\zeta^\star)\Big] - \mathbb{E}\Big[\psi(s,a,s';\widehat{q},w^\star,\widehat{\beta},\zeta^\star)\Big]\right|. && (\varepsilon^B)
\end{aligned}$$

Bounding ε^A: Error from the incorrect indicator ζ.

$$\begin{aligned}
\varepsilon_A &= \gamma\lambda\tau^{-1}\mathbb{E}\widehat{w}(s,a)\Big(\widehat{v}(s') - \widehat{\beta}(s,a)\Big)\Big(\mathbb{I}\Big[\widehat{v}(s') - \widehat{\beta}(s,a) \le 0\Big] - \mathbb{I}\left[v^\star(s') - \beta^\star(s,a) \le 0\right]\Big)\\
&\le C\gamma\lambda\tau^{-1}\mathbb{E}\Big(\widehat{v}(s') - \widehat{\beta}(s,a)\Big)\Big(\mathbb{I}\Big[\widehat{v}(s') - \widehat{\beta}(s,a) \le 0\Big] - \mathbb{I}\left[v^\star(s') - \beta^\star(s,a) \le 0\right]\Big)\\
& && \text{(Assumption 4.2)}\\
&\lesssim \mathbb{E}\Big(\widehat{v}(s') - \widehat{\beta}(s,a)\Big)\Big(\mathbb{I}\Big[\widehat{v}(s') - \widehat{\beta}(s,a) \le 0\Big] - \mathbb{I}\left[v^\star(s') - \beta^\star(s,a) \le 0\right]\Big)
\end{aligned}$$

We break these terms down as follows:

$$\begin{aligned}
&\mathbb{E}\Big(\widehat{v}(s') - \widehat{\beta}(s,a)\Big)\Big(\mathbb{I}\Big[\widehat{v}(s') - \widehat{\beta}(s,a) \le 0\Big] - \mathbb{I}\left[v^\star(s') - \beta^\star(s,a) \le 0\right]\Big)\\
=&\mathbb{E}(v^\star(s') - \beta^\star(s,a))\Big(\mathbb{I}\Big[\widehat{v}(s') - \widehat{\beta}(s,a) \le 0\Big] - \mathbb{I}\left[v^\star(s') - \beta^\star(s,a) \le 0\right]\Big) && (\varepsilon_1^A)\\
&+ \mathbb{E}\Big(\widehat{v}(s') - \widehat{\beta}(s,a) - v^\star(s') + \beta^\star(s,a)\Big)\Big(\mathbb{I}\Big[\widehat{v}(s') - \widehat{\beta}(s,a) \le 0\Big] - \mathbb{I}\left[v^\star(s') - \beta^\star(s,a) \le 0\right]\Big).\\
& && (\varepsilon_2^A)
\end{aligned}$$

We first bound ε_1^A. Assumption 4.2 implies

$$P(0 < |v^\star(s') - \beta^\star(s,a)| \le t) \le c''t,\ \forall t \in [0,c'), \quad P(|v^\star(s') - \beta^\star(s,a)| = 0) = 0,$$

where $c' < 1$ is the min of 1 and the given neighborhood of zero and $c'' \ge 1$ is the max of 1 and the bound on the density in that neighborhood. This implies a margin condition with $\alpha = 1$ and $c = c''/c'$.

We can instantiate the first part of Lemma E.3 with $f(X) = v^\star(s') - \beta^\star(s,a)$, $g(X) = \widehat{v}(s') - \widehat{\beta}(s,a)$ and obtain

$$\begin{aligned}
\varepsilon_1^A &\lesssim \left\|v^\star(s') - \beta^\star(s,a) - \widehat{v}(s') + \widehat{\beta}(s,a)\right\|_p^{\frac{2p}{p+1}}\\
&\le \|\widehat{v}(s') - v^\star(s')\|_p^{\frac{2p}{p+1}} + \left\|\widehat{\beta}(s,a) - \beta^\star(s,a)\right\|_p^{\frac{2p}{p+1}}.
\end{aligned}$$

To bound ε_2^A, first write

$$\begin{aligned}
&\left|\mathbb{E}\Big(\widehat{v}(s') - \widehat{\beta}(s,a) - v^\star(s') + \beta^\star(s,a)\Big)\Big(\mathbb{I}\left[\widehat{v}(s') - \widehat{\beta}(s,a) \le 0\right] - \mathbb{I}\left[v^\star(s') - \beta^\star(s,a) \le 0\right]\Big)\right| \\
&\le \left\|\widehat{v}(s') - \widehat{\beta}(s,a) - v^\star(s') + \beta^\star(s,a)\right\|_p \\
&\quad \cdot \mathbb{P}\Big(\mathbb{I}\left[\widehat{v}(s') - \widehat{\beta}(s,a) \le 0\right] \ne \mathbb{I}\left[v^\star(s') - \beta^\star(s,a) \le 0\right]\Big)^{(p-1)/p}. \qquad \text{(Holder's inequality)}
\end{aligned}$$

We can bound $\mathbb{P}\Big(\mathbb{I}\left[\widehat{v}(s') - \widehat{\beta}(s,a) \le 0\right] \ne \mathbb{I}\left[v^\star(s') - \beta^\star(s,a) \le 0\right]\Big)$ using the second part of Lemma E.3 such that

$$\begin{aligned}
\varepsilon_2^A &\lesssim \left\|\widehat{v}(s') - \widehat{\beta}(s,a) - v^\star(s') + \beta^\star(s,a)\right\|_p \left\|\widehat{v}(s') - \widehat{\beta}(s,a) - v^\star(s') + \beta^\star(s,a)\right\|^{\frac{p-1}{p+1}} \\
&= \left\|\widehat{v}(s') - \widehat{\beta}(s,a) - v^\star(s') + \beta^\star(s,a)\right\|_p^{\frac{2p}{p+1}} \\
&\le \|\widehat{v}(s') - v^\star(s')\|_p^{\frac{2p}{p+1}} + \left\|\widehat{\beta}(s,a) - \beta^\star(s,a)\right\|_p^{\frac{2p}{p+1}}.
\end{aligned}$$

Putting the ε_1^A and ε_2^A together, we have

$$\begin{aligned}
\varepsilon_A &\lesssim \|\widehat{v}(s') - v^\star(s')\|_p^{\frac{2p}{p+1}} + \left\|\widehat{\beta}(s,a) - \beta^\star(s,a)\right\|_p^{\frac{2p}{p+1}} && \text{(when } p \in [1,\infty)) \\
&\lesssim \|\widehat{v}(s') - v^\star(s')\|_\infty^2 + \left\|\widehat{\beta}(s,a) - \beta^\star(s,a)\right\|_\infty^2. && \text{(when } p = \infty)
\end{aligned}$$

Bounding ε^B: Error with correct indicator but wrong nuisances. Now we focus on bounding ε^B.

$$\begin{aligned}
\varepsilon_B &= \mathbb{E}\Big[\psi(s,a,s';\widehat{q},\widehat{w},\widehat{\beta},\zeta^\star)\Big] - \mathbb{E}\Big[\psi(s,a,s';\widehat{q},w^\star,\widehat{\beta},\zeta^\star)\Big] \\
&= \mathbb{E}(\widehat{w}(s,a) - w^\star(s,a))\Big(r(s,a) + \gamma\rho(s,a,s';\widehat{v},\widehat{\beta},\zeta^\star) - \widehat{q}(s,a)\Big) \\
&= \mathbb{E}(\widehat{w}(s,a) - w^\star(s,a))\Big(r(s,a) + \gamma\rho(s,a,s';\widehat{v},\widehat{\beta},\zeta^\star) - \widehat{q}(s,a)\Big) \\
&\quad - \mathbb{E}(\widehat{w}(s,a) - w^\star(s,a))(r(s,a) + \gamma\rho(s,a,s';v^\star,\beta^\star) - q^\star(s,a)) && \text{(Lemma E.4)} \\
&= \mathbb{E}(\widehat{w}(s,a) - w^\star(s,a))\Big(\widehat{q}(s,a) - q^\star(s,a) + \gamma(\rho(s,a,s';\widehat{v},\widehat{\beta},\zeta^\star) - \rho(s,a,s';v^\star,\beta^\star))\Big).
\end{aligned}$$

In the Lemma E.4 step, we used

$0 = (1-\gamma)\mathbb{E}_{d_1} v^\star(s_1) - \mathbb{E}[\psi(s,a,s';q^\star,\widehat{w},\beta^\star)] = (1-\gamma)\mathbb{E}_{d_1} v^\star(s_1) - \mathbb{E}[\psi(s,a,s';q^\star,w^\star,\beta^\star)]$.

Finally, note that

$$\begin{aligned}
&\rho(s,a,s';\widehat{v},\widehat{\beta},\zeta^\star) - \rho(s,a,s';v^\star,\beta^\star) \\
&= (1-\lambda)(\widehat{v}(s') - v^\star(s')) + \lambda\tau^{-1}(\widehat{v}(s') - v^\star(s'))\mathbb{I}\left[\zeta^\star(s,a,s') \le 0\right] \\
&\quad + \lambda(\widehat{\beta}(s,a) - \beta^\star(s,a))\big(1 - \tau^{-1}\mathbb{I}\left[\zeta^\star(s,a,s') \le 0\right]\big).
\end{aligned}$$

Due to continuity of the CDF of $v^\star(s')$ at $\beta^\star(s,a)$ for all s,a, we have $\Pr(\zeta^\star(s',s,a) \le 0 \mid s,a) = \tau$ and so the last term vanishes. Thus, we're left with a quantity that is at most $\lesssim (\widehat{v}(s') - v^\star(s'))$. Therefore,

$$\begin{aligned}
\varepsilon_B &\lesssim \mathbb{E}(\widehat{w}(s,a) - w^\star(s,a))(\mathcal{J}_{U^\pm}(\widehat{q}(s,a) - q^\star(s,a))) \\
&\le \|\mathcal{J}'_{U^\pm}(\widehat{w} - w^\star)\|_2 \|\widehat{q} - q^\star\|_2. && \text{(Holder's inequality)}
\end{aligned}$$

Putting everything together, we obtain the desired rates:

$$\begin{aligned}
|\widehat{V}_{d_1} - V_{d_1}^*| &\lesssim O_p(n^{-1/2}) + \|\mathcal{J}'_{U^\pm}(\widehat{w} - w^\star)\|_2 \|\widehat{q} - q^\star\|_2 + \|\widehat{v} - v^\star\|_p^{\frac{2p}{p+1}} + \left\|\widehat{\beta} - \beta^\star\right\|_p^{\frac{2p}{p+1}} \\
&= O_p(n^{-1/2}) + O_p\Big(r_n^w r_n^q + (r_{n,p}^q)^{\frac{2p}{p+1}} + (r_{n,p}^\beta)^{\frac{2p}{p+1}}\Big) && \text{(when } p \in [1,\infty)) \\
&\lesssim O_p(n^{-1/2}) + \|\mathcal{J}'_{U^\pm}(\widehat{w} - w^\star)\|_2 \|\widehat{q} - q^\star\|_2 + \|\widehat{v} - v^\star\|_\infty^2 + \left\|\widehat{\beta} - \beta^\star\right\|_\infty^2 \\
&= O_p(n^{-1/2}) + O_p\big(r_n^w r_n^q + (r_{n,\infty}^q)^2 + (r_{n,\infty}^\beta)^2\big). && \text{(when } p = \infty)
\end{aligned}$$

E.5 Proof of Normality & Efficiency

In this part of the theorem, we let:

$$\widetilde{V}_{d_1} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_k[\psi(s,a,s';\eta^*)]$$

Then, we can write the following equality:

$$\sqrt{n}(\widehat{V}_{d_1} - V^*_{d_1}) = \sqrt{n}(\widehat{V}_{d_1} - \widetilde{V}_{d_1}) + \underbrace{\sqrt{n}(\widetilde{V}_{d_1} - V^*_{d_1})}_{\xrightarrow{d}\mathcal{N}(0,\Sigma)}$$

The second term converges in distribution to $\mathcal{N}(0,\Sigma)$ from the CLT and the fact that ψ is the efficient influence function. Thus, it remains to show that the first term is $o_p(1)$. We decompose the first term as follows:

$$\sqrt{n}(\widehat{V}_{d_1} - \widetilde{V}_{d_1}) = \sqrt{n}\frac{1}{K}\sum_{k=1}^{n}\Big(\mathbb{E}[\psi(s,a,s';\widehat{\eta}^{[k]})] - \mathbb{E}[\psi(s,a,s';\eta^*)]\Big) \tag{12}$$

$$+ \sqrt{n}\frac{1}{K}\sum_{k=1}^{n}\underbrace{(\mathbb{E}_k - \mathbb{E})[\psi(s,a,s';\widehat{\eta}^{[k]}) - \psi(s,a,s';\eta^*)]}_{\varepsilon_k} \tag{13}$$

In Eq. (12), we have that $|\mathbb{E}[\psi(s,a,s';\widehat{\eta}^{[k]})] - \mathbb{E}[\psi(s,a,s';\eta^*)]|$ is bounded as in Eq. (Rates). Given the theorem's assumption about the nuisance rates, this term is $o_p(n^{-1/2})$ and Eq. (12) is $o_p(1)$. We now seek to control the ε_k term in Eq. (13). Letting $\mathcal{D}_k$ represent the samples in the k^{th} fold, we leverage sample splitting to show that the mean of $\varepsilon_k \mid \mathcal{D}_k$ is 0:

$$\begin{aligned}\mathbb{E}[\varepsilon_k \mid \mathcal{D}_k] &= \mathbb{E}[\mathbb{E}_k[\psi(s,a,s';\widehat{\eta}^{[k]}) - \psi(s,a,s';\eta^*)] - \mathbb{E}[\psi(s,a,s';\widehat{\eta}^{[k]}) - \psi(s,a,s';\eta^*)] \mid \mathcal{D}_k] \\ &= 0\end{aligned}$$

where we consider $\widehat{\eta}^{[k]}$ fixed with respect to the second expectation. The result follows from the fact that $\widehat{\eta}^{[k]}$ does not depend on $\mathcal{D}_k$. Then, we can invoke Chebyshev's inequality to obtain the following bound:

$$P\left(\frac{\varepsilon_k}{\text{Var}[\varepsilon_k \mid \mathcal{D}_k]^{1/2}} \geq \epsilon \middle| \mathcal{D}_k\right) \leq \frac{1}{\epsilon^2},\ \forall \epsilon > 0$$

Thus, $\varepsilon_k \mid \mathcal{D}_k = O_p(\text{Var}[\varepsilon_k \mid \mathcal{D}_k]^{1/2}) = O_p(n^{-1/2}\mathbb{E}[(\psi(s,a,s';\widehat{\eta}^{[k]}) - \psi(s,a,s';\eta^*))^2 \mid \mathcal{D}_k]^{1/2})$. Here, we leveraged the fact that $n_K = n/K$ where K is a fixed integer that doesn't grow with n and the fact that ε_k has 0 conditional mean. For the remainder of the analysis, we leave the conditioning on $\mathcal{D}_k$ implicit for simplicity. To bound $\mathbb{E}[(\psi(s,a,s';\widehat{\eta}^{[k]}) - \psi(s,a,s';\eta^*))^2 \mid \mathcal{D}_k]^{1/2} = \|\psi(s,a,s';\widehat{\eta}^{[k]}) - \psi(s,a,s';\eta^*)\|_2$, we use similar notation and techniques as in Appendix E.4:

$$\|\psi(s,a,s';\widehat{\eta}^{[k]}) - \psi(s,a,s';\eta^*)\|_2 \leq \|\psi(s,a,s';\widehat{q},\widehat{w},\widehat{\beta}) - \psi(s,a,s';\widehat{q},\widehat{w},\widehat{\beta},\zeta^*)\|_2 \tag{σ_1}$$

$$+ \|\psi(s,a,s';\widehat{q},\widehat{w},\widehat{\beta},\zeta^*) - \psi(s,a,s';q^*,w^*,\beta^*,\zeta^*)\|_2 \tag{σ_2}$$

where we invoked Cauchy-Schwarz for the L_2 norm. We bound σ_2 as follows:

$$\begin{aligned}\sigma_2 &\leq \|\psi(s,a,s';\widehat{q},\widehat{w},\widehat{\beta}) - \psi(s,a,s';q^*,\widehat{w},\widehat{\beta},\zeta^*)\|_2 && (\sigma_{2a})\\ &\quad + \|\psi(s,a,s';q^*,\widehat{w},\widehat{\beta},\zeta^*) - \psi(s,a,s';q^*,\widehat{w},\beta^*,\zeta^*)\|_2 && (\sigma_{2b})\\ &\quad + \|\psi(s,a,s';q^*,\widehat{w},\beta^*,\zeta^*) - \psi(s,a,s';q^*,w^*,\beta^*,\zeta^*)\|_2 && (\sigma_{2c})\\ &\leq \|\widehat{v} - v^*\|_2 + \gamma(1-\lambda)\|\widehat{w}\|_2\|\widehat{v} - v^*\|_2 + \gamma\lambda\tau^{-1}\|\widehat{w}\|_2\|\widehat{v} - v^*\|_2 + \|\widehat{w}\|_2\|\widehat{q} - q^*\|_2 && (\sigma_{2a})\\ &\quad + \gamma\lambda\|\widehat{w}\|_2\|\widehat{\beta} - \beta^*\|_2 + \gamma\lambda\tau^{-1}\|\widehat{w}\|_2\|\widehat{\beta} - \beta^*\|_2 && (\sigma_{2b})\\ &\quad + \|\widehat{w} - w^*\|_2\left(\|r\|_2 + \gamma(1-\lambda)\|v^*\|_2 + \gamma\lambda\|\beta^*\|_2 + \gamma\lambda\tau^{-1}\|v^* - \beta^*\|_2\right) && (\sigma_{2c})\end{aligned}$$

Given our rate assumptions, our boundedness assumptions for $\widehat{w}$, the implicit boundedness of q^*, v^*, w^*, β^*, as well as the ordering of the L_2 and L_∞ norms, σ_2 is $o_p(1)$. We now bound the σ_1 term:

$$\sigma_2 = \gamma\lambda\tau^{-1} \left\| \widehat{w}(s,a)(\widehat{v}(s') - \widehat{\beta}(s,a))(\mathbb{I}[\widehat{v}(s') \leq \widehat{\beta}(s,a)] - \mathbb{I}[v^*(s') \leq \beta^*(s,a)]) \right\|_2$$

There are two cases in which the difference of indicators is non-zero:

$$\begin{cases} \widehat{v}(s') \leq \widehat{\beta}(s,a) \text{ and } v^*(s') > \beta^*(s,a) \Rightarrow \mathbb{I}[\widehat{v}(s') \leq \widehat{\beta}(s,a)] - \mathbb{I}[v^*(s') \leq \beta^*(s,a)] = 1 \\ \widehat{v}(s') > \widehat{\beta}(s,a) \text{ and } v^*(s') \leq \beta^*(s,a) \Rightarrow \mathbb{I}[\widehat{v}(s') \leq \widehat{\beta}(s,a)] - \mathbb{I}[v^*(s') \leq \beta^*(s,a)] = -1 \end{cases}$$

In the first case, $\widehat{v}(s') - \widehat{\beta}(s,a) \leq 0, \beta^*(s,a) - v^*(s') < 0$ and thus

$$|(\widehat{v}(s') - \widehat{\beta}(s,a))(\mathbb{I}[\widehat{v}(s') \leq \widehat{\beta}(s,a)] - \mathbb{I}[v^*(s') \leq \beta^*(s,a)])| \leq |\widehat{v}(s') - \widehat{\beta}(s,a) + \beta^*(s,a) - v^*(s')|.$$

In the second case, $\widehat{v}(s') - \widehat{\beta}(s,a) > 0, \beta^*(s,a) - v^*(s') \leq 0$ and

$$|(\widehat{v}(s') - \widehat{\beta}(s,a))(\mathbb{I}[\widehat{v}(s') \leq \widehat{\beta}(s,a)] - \mathbb{I}[v^*(s') \leq \beta^*(s,a)])| \leq |\widehat{v}(s') - \widehat{\beta}(s,a) + \beta^*(s,a) - v^*(s')|.$$

Going back to σ_1, we have:

$$\begin{aligned} \sigma_2 &\leq \gamma\lambda\tau^{-1} \|\widehat{w}\|_2 \|\widehat{v}(s') - \widehat{\beta}(s,a) + \beta^*(s,a) - v^*(s'))\|_2 \\ &\leq \gamma\lambda\tau^{-1} \|\widehat{w}\|_2 (\|\widehat{v} - v^*\|_2 + \|\widehat{\beta} - \beta^*\|_2) \end{aligned}$$

By out theorem's assumptions, this term is also $o_p(1)$. Putting σ_1 and σ_2 together, we have that $\|\psi(s,a,s';\widehat{\eta}^{[k]}) - \psi(s,a,s';\eta^*)\|_2$ is $o_p(1)$ and $\varepsilon_k \mid \mathcal{D}_k$ is $o_p(n^{-1/2})$. By the bounded convergence theorem, this implies that ε_k is also $o_p(n^{-1/2})$. Then, the term in 13 is $o_p(1)$, which further means that $\sqrt{n}(\widehat{V}_{d_1} - \widetilde{V}_{d_1}) = o_p(1)$. Our proof is now complete.

F Proofs for the Efficient Influence Function

We use the ε-contamination approach of [28] to derive an influence function (IF) for our estimand $V^-_{d_1}$. The proof for $V^+_{d_1}$ follows symmetrically. We note that since our tangent space is the whole space as it factorizes in the trivial way (as in [34, Page 54]), the IF we derive is actually the efficient influence function (EIF).

Let $P(s,a,s')$ denote the data distribution. Consider the ε-contamination $P_\varepsilon(s,a,s') = (1 - \varepsilon)P(s,a,s') + \varepsilon\delta(\bar{s},\bar{a},\bar{s}')$, where $\delta(\bar{z})$ is the dirac delta at $\bar{z}$, *i.e.*, $\delta(\bar{z})$ has infinite mass at $\bar{z}$ and 0 mass elsewhere. Let V^-_ε denote the robust value function under the transition kernel $P_\varepsilon(s' \mid s,a)$. Omitting the ε subscript means $\varepsilon = 0$. The IF of $V^-_{d_1}$ is then given by

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}(1-\gamma)\mathbb{E}_{d_1} V^-_\varepsilon(s_1)|_{\varepsilon=0}.$$

We dedicate the rest of this section towards this goal, which will be obtained in Theorem F.5.

Lemma F.1.

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} P_\varepsilon(s' \mid s,a)|_{\varepsilon=0} = \frac{\delta(\bar{s},\bar{a})}{P(s,a)}(\delta(\bar{s}') - P(s' \mid s,a)).$$

Proof. Use the fact $P_\varepsilon(s' \mid s,a) = \frac{P_\varepsilon(s,a,s')}{P_\varepsilon(s,a)} = \frac{(1-\varepsilon)P(s,a,s') + \varepsilon\delta(\bar{s},\bar{a},\bar{s}')}{(1-\varepsilon)P(s,a) + \varepsilon\delta(\bar{s},\bar{a})}$ and take derivative. □

Lemma F.2 (IF of conditional expectation). *For any s, a and f_ε,*

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathbb{E}_{P_\varepsilon}[f_\varepsilon(s') \mid s,a]|_{\varepsilon=0} = \frac{\delta(\bar{s},\bar{a})}{P(s,a)}(f(\bar{s}') - \mathbb{E}_P[f(s') \mid s,a]) + \mathbb{E}_P\left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} f_\varepsilon(s')|_{\varepsilon=0} \mid s,a\right],$$

where $f = f_0$.

Proof.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathbb{E}_{P_\varepsilon}[f_\varepsilon(s') \mid s,a]|_{\varepsilon=0} &= \sum_{s'} f(s')\frac{\mathrm{d}}{\mathrm{d}\varepsilon}P_\varepsilon(s' \mid s,a)|_{\varepsilon=0} + \sum_{s'}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f_\varepsilon(s')|_{\varepsilon=0}P(s' \mid s,a) \\ &= \frac{\delta(\bar{s},\bar{a})}{P(s,a)}(f_0(\bar{s}') - \mathbb{E}_P[f_0(s') \mid s,a]) + \mathbb{E}_P\left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f_\varepsilon(s')|_{\varepsilon=0} \mid s,a\right],\end{aligned}$$

□

Lemma F.3 (IF of conditional CVaR). *For any τ, s, a and f_ε,*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{CVaR}_{\tau,P_\varepsilon}[f_\varepsilon(s') \mid s,a]|_{\varepsilon=0} = \frac{\delta(\bar{s},\bar{a})}{P(s,a)}\big(&\beta_\tau(s,a) + \tau^{-1}(f(\bar{s}') - \beta_\tau(s,a))_- - \mathrm{CVaR}_\tau(f(s') \mid s,a)\big) \\ &+ \mathbb{E}_P\left[\tau^{-1}\mathbb{I}\left[f(s') \le \beta_\tau(s,a)\right]\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f_\varepsilon(s')|_{\varepsilon=0} \mid s,a\right],\end{aligned}$$

where $f = f_0$ and $\beta_\tau(s,a)$ be the $(1-\tau)$-th quantile of $f(s'), s' \sim P(s,a)$.

Proof.

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathrm{CVaR}_{P_\varepsilon}[f_\varepsilon(s') \mid s,a]|_{\varepsilon=0} \tag{14}$$

$$= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\min_b \mathbb{E}_{P_\varepsilon}\left[b + \tau^{-1}(f_\varepsilon(s') - b)_- \mid s,a\right]|_{\varepsilon=0} \tag{15}$$

$$= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathbb{E}_{P_\varepsilon}\left[\beta_\tau(s,a) + \tau^{-1}(f_\varepsilon(s') - \beta_\tau(s,a))_- \mid s,a\right]|_{\varepsilon=0}, \tag{16}$$

where the last equality is due to Danskin's theorem and the fact that $\beta_\tau(s,a)$ is the maximizer of the CVaR dual form at $\varepsilon = 0$. Continuing, let $g_\varepsilon(s'; s,a) := \beta_\tau(s,a) + \tau^{-1}(f_\varepsilon(s') - \beta_\tau(s,a))_-$, so

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathbb{E}_{P_\varepsilon}[g_\varepsilon(s'; s,a) \mid s,a] \\ &= \frac{\delta(\bar{s},\bar{a})}{P(s,a)}(g(\bar{s}'; s,a) - \mathbb{E}_P[g(s', s,a) \mid s,a]) + \mathbb{E}_P\left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon}g_\varepsilon(s'; s,a)|_{\varepsilon=0} \mid s,a\right] && \text{(Lemma F.2)} \\ &= \frac{\delta(\bar{s},\bar{a})}{P(s,a)}(g(\bar{s}'; s,a) - \mathrm{CVaR}_\tau(f(s') \mid s,a)) + \mathbb{E}_P\left[\tau^{-1}\mathbb{I}\left[f(s') \le \beta_\tau(s,a)\right]\frac{\mathrm{d}}{\mathrm{d}\varepsilon}f_\varepsilon(s')|_{\varepsilon=0} \mid s,a\right].\end{aligned}$$

This concludes the proof. □

We now prove the key "one-step forward" lemma.

Lemma F.4 (One-Step Forward). *For any state distribution $\nu(s)$, we have*

$$\begin{aligned}&\mathbb{E}_{s\sim\nu}\left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon}V^-_\varepsilon(s)|_{\varepsilon=0}\right] \\ &= \frac{\nu(\bar{s})\pi(\bar{a} \mid \bar{s})}{P(\bar{s},\bar{a})}\big(r(\bar{s},\bar{a}) + \gamma\big((1-\lambda)V^-(\bar{s}') + \lambda\big(\beta_\tau(\bar{s},\bar{a}) + \tau^{-1}(V^-(\bar{s}') - \beta_\tau(\bar{s},\bar{a}))_-\big)\big) \\ &\qquad\qquad - Q^-(\bar{s},\bar{a})\big) \\ &\quad + \gamma\mathbb{E}_{s\sim\nu}\left[\mathbb{E}_{\pi,P}\left[\big((1-\lambda) + \lambda\tau^{-1}\mathbb{I}\left[V^-(s') \le \beta_\tau(s,a)\right]\big)\frac{\mathrm{d}}{\mathrm{d}\varepsilon}V^-_\varepsilon(s')|_{\varepsilon=0} \mid s\right]\right].\end{aligned}$$

Proof. For any s_1, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_\varepsilon^-(s_1) \\
&= \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathbb{E}_{a_1 \sim \pi(s_1)} \left[r(s_1, a_1) + \gamma((1-\lambda)\mathbb{E}_{P_\varepsilon}\left[V_\varepsilon^-(s_2) \mid s_1, a_1\right] + \lambda \operatorname{CVaR}_{\tau, P_\varepsilon}\left[V_\varepsilon^-(s_2) \mid s_1, a_1\right]\right]_{\varepsilon=0} \\
&= \gamma \mathbb{E}_{a_1 \sim \pi(s_1)} \left[(1-\lambda) \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathbb{E}_{\tau, P_\varepsilon}\left[V_\varepsilon^-(s_2) \mid s_1, a_1\right]|_{\varepsilon=0} + \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \operatorname{CVaR}_{\tau, P_\varepsilon}\left[V_\varepsilon^-(s_2) \mid s_1, a_1\right]|_{\varepsilon=0} \right] \\
&= \gamma(1-\lambda) \mathbb{E}_{a_1 \sim \pi(s_1)} \left[\frac{\delta(\bar{s}, \bar{a})}{P(s_1, a_1)} \big(V^-(\bar{s}') - \mathbb{E}_P\big[V^-(s_2) \mid s_1, a_1\big] \big) \right] \\
&+ \gamma(1-\lambda) \mathbb{E}_{a_1 \sim \pi(s_1)} \mathbb{E}_P \left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_\varepsilon^-(s_2)|_{\varepsilon=0} \mid s_1, a_1 \right] \\
&+ \gamma\lambda \mathbb{E}_{a_1 \sim \pi(s_1)} \left[\frac{\delta(\bar{s}, \bar{a})}{P(s_1, a_1)} \big(\beta_\tau(s_1, a_1) + \tau^{-1}(V^-(\bar{s}') - \beta_\tau(s_1, a_1))_- - \operatorname{CVaR}_\tau(V^-(s_2) \mid s_1, a_1) \big) \right] \\
&+ \gamma\lambda \mathbb{E}_{a_1 \sim \pi(s_1)} \mathbb{E}_P \left[\tau^{-1} \mathbb{I}\left[V^-(s_2) \leq \beta_\tau(s_1, a_1)\right] \frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_{\pi, P_\varepsilon}^-(s_2) \right].
\end{aligned}$$

Taking expectation over $s_1 \sim \nu$, we have

$$\begin{aligned}
\mathbb{E}_{s \sim \nu}\left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_\varepsilon^-(s)|_{\varepsilon=0} \right] = \gamma \frac{\nu(\bar{s})\pi(\bar{a} \mid \bar{s})}{P(\bar{s}, \bar{a})} \Bigg(&(1-\lambda)V^-(\bar{s}') + \lambda\big(\beta_\tau(\bar{s}, \bar{a}) + \tau^{-1}(V^-(\bar{s}') - \beta_\tau(\bar{s}, \bar{a}))_-\big) \\
&- \big((1-\lambda)\mathbb{E}\big[V^-(s') \mid \bar{s}, \bar{a}\big] + \lambda \operatorname{CVaR}_\tau(V^-(s') \mid \bar{s}, \bar{a})\big) \Bigg) \\
&+ \gamma \mathbb{E}_{s \sim \nu}\left[\mathbb{E}_{\pi, P}\left[\big((1-\lambda) + \lambda\tau^{-1}\mathbb{I}\left[V^-(s') \leq \beta_\tau(s, a)\right]\big) \frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_\varepsilon^-(s')|_{\varepsilon=0} \mid s \right] \right].
\end{aligned}$$

Finally recall that V^- satisfies the Bellman equation, so

$$(1-\lambda)\mathbb{E}\big[V^-(s') \mid \bar{s}, \bar{a}\big] + \lambda \operatorname{CVaR}_\tau(V^-(s') \mid \bar{s}, \bar{a}) = Q^-(\bar{s}, \bar{a}) - r(\bar{s}, \bar{a}).$$

This concludes the proof. □

Equipped with our main one-step lemma, we can now unroll it an infinite number of steps to derive the IF of our estimand.

Theorem F.5 (IF of Estimand)**.** *Let us denote*

$$g(\bar{s}, \bar{a}, \bar{s}') := r(\bar{s}, \bar{a}) + \gamma\big((1-\lambda)V^-(\bar{s}') + \lambda\big(\beta_\tau(\bar{s}, \bar{a}) + \tau^{-1}(V^-(\bar{s}') - \beta_\tau(\bar{s}, \bar{a}))_-\big)\big).$$

Then, we have

$$\mathbb{E}_{d_1}\left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_\varepsilon^-(s_1)|_{\varepsilon=0} \right] = \frac{d_{\mathit{rob}}^{\pi, \infty}(\bar{s}, \bar{a})}{P(\bar{s}, \bar{a})} g(\bar{s}, \bar{a}, \bar{s}').$$

Proof. Let d_h denote the h-th step visitation in the robust MDP, with transition P_{rob} satisfying $\frac{P_{\text{rob}}(s' \mid s, a)}{P(s' \mid s, a)} = (1-\lambda) + \lambda\tau^{-1}\mathbb{I}\left[V^-(s') \leq \beta_\tau(s, a)\right]$. Then notice that the final term of Lemma F.4 is exactly $\mathbb{E}_{s \sim \nu}\big[\mathbb{E}_{\pi, P_{\text{rob}}}\big[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_\varepsilon^-(s')|_{\varepsilon=0} \mid s\big]\big]$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{d_1}\left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_\varepsilon^-(s_1)|_{\varepsilon=0} \right] \\
&= \frac{d_1(\bar{s})\pi(\bar{a} \mid \bar{s})}{P(\bar{s}, \bar{a})} g(\bar{s}, \bar{a}, \bar{s}') + \gamma \mathbb{E}_{s_2 \sim d_2}\left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_\varepsilon^-(s_2)|_{\varepsilon=0} \right] \\
&= \frac{d_1(\bar{s})\pi(\bar{a} \mid \bar{s})}{P(\bar{s}, \bar{a})} g(\bar{s}, \bar{a}, \bar{s}') + \gamma \frac{d_2(\bar{s})\pi(\bar{a} \mid \bar{s})}{P(\bar{s}, \bar{a})} g(\bar{s}, \bar{a}, \bar{s}') + \gamma^2 \mathbb{E}_{s_3 \sim d_3}\left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_\varepsilon^-(s_3)|_{\varepsilon=0} \right].
\end{aligned}$$

Iterating the process, we have

$$\mathbb{E}_{d_1}\left[\frac{\mathrm{d}}{\mathrm{d}\varepsilon} V_\varepsilon^-(s_1)|_{\varepsilon=0} \right] = \sum_{h=1}^{\infty} \gamma^{h-1} \frac{d_h(\bar{s})\pi(\bar{a} \mid \bar{s})}{P(\bar{s}, \bar{a})} g(\bar{s}, \bar{a}, \bar{s}') = \frac{d_{\text{rob}}^{\pi, \infty}(\bar{s}, \bar{a})}{P(\bar{s}, \bar{a})} g(\bar{s}, \bar{a}, \bar{s}'),$$

as desired. □

Finally, we can conclude that the IF in Theorem F.5 is in fact the efficient IF (EIF) because it is in the tangent space, as the tangent space is contains all functions [34].

G Properties of the Robust MDP

G.1 Identification of robust Q

Lemma 3.1. *Set $\tau(s,a) = (\Lambda(s,a)+1)^{-1}$. Then, for any $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,*

$$\mathcal{T}^{\pm}_{\mathsf{rob}} q(s,a) = r(s,a) + \gamma\Lambda^{-1}(s,a)\mathbb{E}[v(s') \mid s,a] + \gamma(1-\Lambda^{-1}(s,a))\,\mathrm{CVaR}^{\pm}_{\tau(s,a)}[v(s') \mid s,a],$$

where $v(s') = \mathbb{E}_{a' \sim \pi_t(s')}[q(s',a')]$, and $\mathbb{E}, \mathrm{CVaR}_\tau$ are under the observed kernel $P(\cdot \mid s,a)$.

Proof. We rewrite Eq. (1) as:

$$0 \le \frac{U(s' \mid s,a) - \Lambda^{-1}(s,a)P(s' \mid s,a)}{P(s' \mid s,a)} \le \Lambda(s,a) - \Lambda^{-1}(s,a).$$

Note that $\Lambda - \Lambda^{-1} = (1-\Lambda^{-1})(1+\Lambda)$. Naming $G(s' \mid s,a) = \frac{U(s'|s,a) - \Lambda^{-1}(s,a)P(s'|s,a)}{1-\Lambda^{-1}(s,a)}$, we have $G(\cdot \mid s,a) \ll P(\cdot \mid s,a)$ and $\|\frac{\mathrm{d}G(s'|s,a)}{\mathrm{d}P(s'|s,a)}\| \le \Lambda(s,a)+1$. Setting $\tau(s,a) = \frac{1}{\Lambda(s,a)+1}$, we can apply the primal form of CVaR to obtain:

$$\sup_{G \ll P : \|\frac{\mathrm{d}G(\cdot|s,a)}{\mathrm{d}P(\cdot|s,a)}\|_\infty \le \tau^{-1}(s,a)} \qquad = \mathrm{CVaR}^{+}_{\tau(s,a)}[f(s') \mid s,a],$$

where the CVaR is over P [3]. The other case, with $\inf$ and lower CVaR, is identical. □

G.2 Identification of robust kernel and visitation

Lemma 4.1. *Suppose $F^{\pm}(\beta^{\pm}_\tau(s,a) \mid s,a) = \frac{1}{2} \pm (\frac{1}{2} - \tau)$, where $\beta^{\pm}_\tau(s,a)$ is the upper/lower τ-th quantile of $F^{\pm}(\cdot \mid s,a)$. Then,*

$$U^{\pm}(s' \mid s,a)/P(s' \mid s,a) = \Lambda^{-1}(s,a) + (1-\Lambda^{-1})\tau(s,a)^{-1}\mathbb{I}[\pm(V^{\pm}(s') - \beta^{\pm}_\tau(s,a)) \ge 0]. \quad (4)$$

Lemma G.1. *Fix any $v : \mathcal{S} \to \mathbb{R}$ and define the pushforward $F_v(y \mid s,a) = P(v(s') \le y \mid s,a)$. Suppose $F_v(\beta^{\pm}_{\tau,F_v(\cdot|s,a)}(s,a) \mid s,a) = \frac{1}{2} \pm (\frac{1}{2} - \tau)$, where $\beta^{\pm}_{\tau,F_v}$ is the upper/lower τ-quantile of F_v. Then, $\sup_{U \in \mathcal{U}(P)} \mathbb{E}_U[v(s') \mid s,a] = \mathbb{E}_{s' \sim U^{+}_v(s,a)}[v(s')]$ and $\inf_{U \in \mathcal{U}(P)} \mathbb{E}_U[v(s') \mid s,a] = \mathbb{E}_{s' \sim U^{-}_v(s,a)}[v(s')]$, where*

$$U^{\pm}_v(s' \mid s,a)/P(s' \mid s,a) = \Lambda^{-1}(s,a) + (1-\Lambda^{-1})\tau(s,a)^{-1}\mathbb{I}[\pm(v(s') - \beta^{\pm}_{\tau,F_v(\cdot|s,a)}(s,a)) \ge 0].$$

Proof. Recall that $\mathrm{CVaR}^{+}_\tau(v(s') \mid s,a) = \mathbb{E}[v(s') \mid f(s') \ge \beta^{+}_\tau(s,a), s, a]$ under the assumption that the CDF of $v(s')$ is differentiable at $\beta^{+}_\tau(s,a)$. Then the result follows immediately from Lemma 3.1 by noticing that the form of U^{+} exactly recovers the convex combination of expectation and CVaR. Alternatively, one can use the closed form solution of the primal CVaR to obtain the result, as in [3]. The proof for the other case, with $\inf$ and lower CVaR, is identical. □

H Proofs for Robust FQE

We prove a more general result with approximate completeness, which shows that Theorem 3.4 is robust to approximate completeness.

Assumption H.1 (Approximate Completeness). $\max_{q \in \mathcal{Q}} \min_{g \in \mathcal{Q}} \|g - \mathcal{T}^{\pm}_{\mathrm{CVaR}} q\|_\nu \le \varepsilon_{\mathsf{QComp}}$.

Theorem H.2. *Assume Assumption H.1. Under the same setup as Theorem 3.4, we have*

$$\|\widehat{q}_K^{\pm} - Q^{\pm}\|_{\mu} \lesssim \frac{1}{(1-\gamma)^2}(\sqrt{C_{\mu}^{\pm}} \cdot (\varepsilon_n^{\mathcal{Q}} + \varepsilon_{\mathsf{QComp}}) + \mathrm{err}_{\mathsf{QR}}^2(n/2K, \delta/2K)),$$

and

$$\left|V_{d_1}^{\pm} - (1-\gamma)\mathbb{E}_{d_1}[\widehat{q}_K^{\pm}(s_1, \pi_t)]\right| \lesssim \gamma^K + \frac{1}{1-\gamma}(\sqrt{C_{\mu}^{\pm}} \cdot (\varepsilon_n^{\mathcal{Q}} + \varepsilon_{\mathsf{QComp}}) + \mathrm{err}_{\mathsf{QR}}^2(n/2K, \delta/2K)).$$

Proof. Let $U^{\pm}$ denote the worst-case kernel that satisfies $V_{d_1}^{\pm} = (1-\gamma)\mathbb{E}_{d_1} V_{U^{\pm}}^{\pi_{\mathsf{t}}}(s_1)$. Then,

$$\begin{aligned} V_{d_1}^{\pm} - (1-\gamma)\mathbb{E}_{d_1}[\widehat{q}_K^{\pm}(s_1, \pi_{\mathsf{t}})] &= (1-\gamma)\mathbb{E}_{d_1}[V_{U^{\pm}}^{\pi_{\mathsf{t}}}(s_1) - \widehat{q}_K(s_1, \pi_{\mathsf{t}})] \\ &= \mathbb{E}_{d_{U^{\pm}}^{\pi,\infty}}[\mathcal{T}_{U^{\pm}}^{\pi_{\mathsf{t}}} \widehat{q}_K(s,a) - \widehat{q}_K(s,a)] && \text{(Lemma H.3)} \\ &\leq \frac{4}{1-\gamma} \max_{k=1,2,\dots} \|\widehat{q}_k - \mathcal{T}_{U^{\pm}}^{\pi_{\mathsf{t}}} \widehat{q}_{k-1}\|_{d_{U^{\pm}}^{\pi_{\mathsf{t}},\infty}} + \gamma^{K/2}. && \text{(Lemma H.4)} \end{aligned}$$

Consider any $k = 1, 2, \dots$. By definition of $U^{\pm}$, we have

$$\|\widehat{q}_k - \mathcal{T}_{U^{\pm}}^{\pi_{\mathsf{t}}} \widehat{q}_{k-1}\|_{d_{U^{\pm}}^{\pi_{\mathsf{t}},\infty}} = \left\|\widehat{q}_k - \mathcal{T}_{\beta_k^{\star}}^{\pm} \widehat{q}_{k-1}\right\|_{d^{\pm,\infty}}, \qquad \text{(by def of } U^{\pm})$$

where $\beta_k^{\star}(s,a)$ is the true quantile of $\widehat{v}_{k-1}(s')$. Denote $q_k^{\star} := \mathcal{T}_{\mathsf{rob}}^{\pm} \widehat{q}_{k-1}$ and let $\beta_k^{\star}$ be the true upper/lower quantile of $\widehat{q}_{k-1}$. Recall the population loss function is

$$\begin{aligned} L_k(q, \beta) &:= \mathbb{E}\left[\left(y_k^{\beta}(s,a,s') - q(s,a)\right)^2\right] \\ y_k^{\beta}(s,a,s') &= r(s,a) + \gamma \Lambda^{-1}(s,a)\widehat{v}_{k-1}(s') \\ &\quad + \gamma(1 - \Lambda^{-1}(s,a))\big(\beta(s,a) + \tau^{-1}(s,a)(\widehat{v}_{k-1}(s') - \beta(s,a))_{\pm}\big). \end{aligned}$$

The empirical loss $\widehat{L}_k(q, \beta)$ is if $\mathbb{E}$ is replaced by $\mathbb{E}_n$. Note that $\widehat{q}_k = \arg\min_{q \in \mathcal{Q}} \widehat{L}_k(q, \widehat{\beta}_k)$.

Nonparametric Least Squares with Model Misspecification. We will directly invoke [62, Theorem 13.13], which gives a fast rate for misspecified least squares with general nonparametric classes. We now bound the misspecification. Recall that at the k-th iteration, our regression Bayes-optimal is $\mathbb{E}[y_k^{\widehat{\beta}_k}(s,a,s') \mid s,a] = \mathcal{T}_{\widehat{\beta}_k} \widehat{q}_{k-1}(s,a)$. By Lemma E.3, we know this is close to $\mathcal{T}_{\beta_k^{\star}} \widehat{q}_{k-1}(s,a)$ with second order errors in β: for any μ, we have

$$\left\|\mathcal{T}_{\widehat{\beta}_k}^{\pm} \widehat{q}_{k-1} - \mathcal{T}_{\beta_k^{\star}}^{\pm} \widehat{q}_{k-1}\right\|_{d_{\mu}^{\pm,\infty}} \lesssim \|\widehat{\beta}_k - \beta_k^{\star}\|_{\infty}^2.$$

Finally, by approximate completeness (Assumption H.1), there exists $g \in \mathcal{Q}$ such that $\|\mathcal{T}_{\beta_k^{\star}} \widehat{q}_{k-1}(s,a) - g\| \leq \varepsilon_{\mathsf{QComp}}$. Putting this together: for any k, there exists a $g \in \mathcal{Q}$ such that

$$\begin{aligned} \|g - \mathcal{T}_{\widehat{\beta}_k} \widehat{q}_{k-1}(s,a)\|_{d_{\mu}^{\pm,\infty}} &\leq \|g - \mathcal{T}_{\beta_k^{\star}} \widehat{q}_{k-1}(s,a)\|_{d_{\mu}^{\pm,\infty}} + \|\mathcal{T}_{\beta_k^{\star}} \widehat{q}_{k-1}(s,a) - \mathcal{T}_{\widehat{\beta}_k} \widehat{q}_{k-1}(s,a)\|_{d_{\mu}^{\pm,\infty}} \\ &\leq \sqrt{C_{\mu}^{\pm}} \cdot \varepsilon_{\mathsf{QComp}} + \|\widehat{\beta}_k - \beta_k^{\star}\|_{\infty}^2. \end{aligned}$$

Therefore, [62, Theorem 13.13] (and concentration of least squares) certifies that:

$$\left\|\widehat{q}_k - \mathcal{T}_{\widehat{\beta}_k} \widehat{q}_{k-1}\right\|_{d^{\pm,\infty}} \lesssim \sqrt{C_{\mu}^{\pm}} \cdot (\varepsilon_{\mathsf{QComp}} + \varepsilon_n) + \|\widehat{\beta}_k - \beta_k^{\star}\|_{\infty}^2.$$

Therefore, we have proven:

$$\begin{aligned} \left\|\widehat{q}_k - \mathcal{T}_{\beta_k^{\star}}^{\pm} \widehat{q}_{k-1}\right\|_{d_{\mu}^{\pm,\infty}} &\leq \left\|\widehat{q}_k - \mathcal{T}_{\widehat{\beta}_k}^{\pm} \widehat{q}_{k-1}\right\|_{d_{\mu}^{\pm,\infty}} + \left\|\mathcal{T}_{\widehat{\beta}_k}^{\pm} \widehat{q}_{k-1} - \mathcal{T}_{\beta_k^{\star}}^{\pm} \widehat{q}_{k-1}\right\|_{d_{\mu}^{\pm,\infty}} \\ &\lesssim \sqrt{C_{\mu}^{\pm}} \cdot (\varepsilon_{\mathsf{QComp}} + \varepsilon_n) + \|\widehat{\beta}_k - \beta_k^{\star}\|_{\infty}^2. \end{aligned}$$

This concludes the proof.

□

Lemma H.3 (Performance Difference)**.** *For any π, transition kernel P, and function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we have*

$$V_P^\pi - \mathbb{E}_{s\sim d_1}[f(s,\pi)] = \frac{1}{1-\gamma}\mathbb{E}_{d_P^{\pi,\infty}}[\mathcal{T}_P^\pi f(s,a) - f(s,a)].$$

Proof. See Lemma C.1 of [14]. □

Lemma H.4 (Unrolling)**.** *For any π, transition kernel P, and functions $f_0, f_1, \dots, f_K : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ satisfying $f_0(s,a) = 0$, we have $\|f_K - \mathcal{T}_P^\pi f_K\|_{d_P^{\pi,\infty}} \le \frac{4}{1-\gamma}\max_{k=1,2,\dots}\|f_k - \mathcal{T}_P^\pi f_{k-1}\|_{d_P^{\pi,\infty}} + \gamma^{K/2}$.*

Proof. See Lemma C.2 of [14]. □

I Proofs for Robust Minimax Algorithm

Assumption I.1 (Approximate W-realizability and completeness)**.** Assume the following hold for $\mathcal{W}$ and $\mathcal{F}$:
(A) Approximate realizability: $\min_{w\in\mathcal{W}}\|\mathcal{J}_{U^\pm}(w^\pm - w)\|_2 \le \varepsilon_{\mathsf{WReal}}$;
(B) Approximate completeness: $\max_{w\in\mathcal{W}}\min_{f\in\mathcal{F}}\left\|f - \mathcal{J}'_{U^\pm}(w - w^\pm)\right\|_2 \le \varepsilon_{\mathsf{WComp}}$.

We prove a more general result with approximate realizability and completeness, which implies Theorem 4.4 that is robust to misspecification in its assumptions.

Theorem I.2. *Under Assumption I.1 and the same setup as Theorem 4.4, we have*

$$\left\|\mathcal{J}'_{U^\pm}(\widehat{w} - w^\pm)\right\|_2 \lesssim \varepsilon_n^{\mathcal{W}} + \|\widetilde{\zeta}^\pm - \zeta^\pm\|_\infty + \sqrt{\frac{\log(1/\delta)}{n}} + \varepsilon_{\mathsf{WReal}} + \varepsilon_{\mathsf{WComp}}.$$

Proof. For this proof, we focus on the worst-case kernel $P^\star$ of the form $\frac{P^\star(s'|s,a)}{P(s'|s,a)} = \tau^{-1}(s,a)\mathbb{I}[\zeta^\star(s,a,s') \le 0]$ where $\zeta^\star(s,a,s') = V^-(s') - \beta^-(s,a)$. This corresponds to the pure CVaR case of $\mathcal{T}_{\text{rob}}^-$; the $\mathbb{E}$ part is identical to standard non-robust RL so we omit it. The best-case kernel U^+ can be handled similarly. Let $\widehat{P}(s' \mid s,a)$ denote our estimated robust kernel, which satisfies $\frac{\widehat{P}(s'|s,a)}{P(s'|s,a)} = \tau^{-1}(s,a)\mathbb{I}[\widehat{\zeta}(s,a,s') \le 0]$, where $\widehat{\zeta}(s,a,s')$ is the given prior stage estimate of $\zeta^\star(s,a,s') = V^-(s') - \beta^-(s,a)$.

The key and only difference between our Algorithm 2 and the MIL algorithm ($\widehat{w}_{\text{mil}}$) of [59] is that our next-state samples are importance weighted with $\xi^\pm(s,a,s')$, which is the density ratio of the estimated robust kernel $\widehat{P}(s' \mid s,a)$ and the nominal kernel $P(s' \mid s,a)$. Note also that $\xi^\pm(s,a,s') \le \tau^{-1}(s,a) < \infty$, and hence $|\mathbb{E}_n[\zeta(s,a,s')f(s')] - \mathbb{E}_{s,a\sim\nu,s'\sim\widehat{P}(s,a)}[f(s')]| \lesssim \sqrt{\log(1/\delta)/n}$ w.p. $1-\delta$. Therefore, up to $\mathcal{O}(\sqrt{\log(1/\delta)/n})$ errors, our Algorithm 2 can be viewed as MIL applied to the MDP with kernel $\widehat{P}$.

To invoke the result of [59, Theorem 6.1] (in MDP with kernel $\widehat{P}$), we need to show that its assumptions are met by bounding the model misspecification, *i.e.*, Eq. (6) and Appendix C of [59]. Note that these misspecifications are w.r.t. the MDP with kernel $\widehat{P}$, since this is the MDP in which we're applying Theorem 6.1 of [59]. Specifically, the two errors we need to bound are, (A) approximate realizability: $\varepsilon_A = \min_{w\in\mathcal{W}}\|\mathcal{J}'_{\widehat{P}}(w_{\widehat{P}} - w)\|_2$; and (B) approximate completeness: $\varepsilon_B = \max_{w\in\mathcal{W}}\min_{f\in\mathcal{F}}\|f - \mathcal{J}'_{\widehat{P}}(w - w_{\widehat{P}})\|_2$ where recall that $\mathcal{J}_P$ is the linear operator defined as $\mathcal{J}_P f(s,a) := \gamma\mathbb{E}_P[f(s',\pi_{\text{t}}) \mid s,a] - f(s,a)$ and $\mathcal{J}'_P$ is the adjoint.

Bounding misspecifications by $\|\widehat{\zeta} - \zeta^\star\|_\infty$. Since $\zeta^\star(s,a,s')$ has a marginal CDF that's boundedly differentiable around 0 (*i.e.*, (ii) of Assumption 4.2), [32, Lemma 3] implies that $\zeta^\star(s,a,s')$ satisfies a 1-margin (Definition E.2). Hence, Lemma E.3 and the continuity of $\zeta^\star(s,a,s')$ implies that

$$\begin{aligned}
&\Pr\Big(\mathbb{I}[\widehat{\zeta}(s,a,s') \le 0] \neq \mathbb{I}[\zeta^\star(s,a,s') \le 0]\Big) \\
&= \Pr\Big((\mathbb{I}[\widehat{\zeta}(s,a,s') \le 0] \neq \mathbb{I}[\zeta^\star(s,a,s') \le 0]), \zeta^\star(s,a,s') \neq 0\Big) \lesssim \|\widehat{\zeta} - \zeta^\star\|_\infty,
\end{aligned}$$

Thus, for any $v : \mathcal{S} \to \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\big|(\mathbb{E}_{\widehat{P}} - \mathbb{E}_{P^\star})[v(s') \mid s,a]\big| &\le \mathbb{E}[\tau^{-1}(s,a)(\mathbb{I}[\widehat{\zeta}(s,a,s') \le 0] \neq \mathbb{I}[\zeta^\star(s,a,s') \le 0]) \cdot |v(s')|] \\
&\lesssim \|v\|_\infty \cdot \Pr\Big(\mathbb{I}[\widehat{\zeta}(s,a,s') \le 0] \neq \mathbb{I}[\zeta^\star(s,a,s') \le 0]\Big) \\
&\lesssim \|v\|_\infty \|\widehat{\zeta} - \zeta^\star\|_\infty,
\end{aligned}$$

or equivalently

$$\mathbb{E}\|\widehat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_{\mathsf{TV}} \lesssim \|\widehat{\zeta} - \zeta^\star\|_\infty. \tag{17}$$

Equipped with Eq. (17), we can now bound the following two types of errors: (i) $\langle f, (\mathcal{T}_{P^\star} - \mathcal{T}_{\widehat{P}})g\rangle$, and (ii) $\langle w_{\widehat{P}} - w_{P^\star}, h\rangle$, where $f, g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $h : \mathcal{S} \to \mathbb{R}$, and $\mathcal{T}_P$ and w_P are the Bellman operator and visitation density of target policy π_{t} in the MDP with kernel P.

For (i):

$$\begin{aligned}
\big|\langle f, (\mathcal{J}_{P^\star} - \mathcal{J}_{\widehat{P}})g\rangle\big| &= \big|\mathbb{E}[f(s,a)\big(\gamma(\mathbb{E}_{P^\star} - \mathbb{E}_{\widehat{P}})[g(s', \pi_{\mathsf{t}}) \mid s,a]\big)]\big| \\
&\le \gamma \|f\|_\infty \mathbb{E}\big|(\mathbb{E}_{P^\star} - \mathbb{E}_{\widehat{P}})[g(s', \pi_{\mathsf{t}}) \mid s,a]\big| \\
&\lesssim \gamma \|f\|_\infty \|g(\cdot, \pi_{\mathsf{t}})\|_\infty \|\widehat{\zeta} - \zeta^\star\|_\infty.
\end{aligned}$$

For (ii):

$$\begin{aligned}
\langle w_{\widehat{P}} - w_{P^\star}, h\rangle &= \mathbb{E}[(w_{\widehat{P}}(s) - w_{P^\star}(s))h(s)] \\
&\le \|h\|_\infty \|d_{\widehat{P}} - d_{P^\star}\|_{\mathsf{TV}} \\
&\le \|h\|_\infty \frac{\gamma}{1-\gamma} \mathbb{E}_{d_{P^\star}} \|\widehat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_{\mathsf{TV}} && \text{(Eq. (19))} \\
&\lesssim C\|h\|_\infty \frac{\gamma}{1-\gamma} \mathbb{E}\|\widehat{P}(\cdot \mid s,a) - P^\star(\cdot \mid s,a)\|_{\mathsf{TV}} && \text{(Assumption 4.2(i))} \\
&\lesssim C\|h\|_\infty \frac{\gamma}{1-\gamma} \|\widehat{\zeta} - \zeta^\star\|_\infty,
\end{aligned}$$

where $C = \|{}^{\mathrm{d}d^{P^\star}}\!/{}_{\mathrm{d}\nu}\|_\infty < \infty$.

For approximate realizability (ε_A): for any $w \in \mathcal{W}$, we have

$$\begin{aligned}
&\|\mathcal{J}'_{\widehat{P}}(w_{\widehat{P}} - w)\|_2 \\
&\le \|(\mathcal{J}_{\widehat{P}} - \mathcal{J}_{P^\star})'(w_{\widehat{P}} - w)\|_2 + \|\mathcal{J}'_{P^\star}(w_{\widehat{P}} - w_{P^\star})\|_2 + \|\mathcal{J}'_{P^\star}(w^\star - w)\|_2 \\
&= \langle w_{\widehat{P}} - w, (\mathcal{J}_{\widehat{P}} - \mathcal{J}_{P^\star})g_1\rangle + \langle w_{\widehat{P}} - w_{P^\star}, \mathcal{J}_{P^\star} g_2\rangle + \|\mathcal{J}'_{P^\star}(w^\star - w)\|_2 \\
&\lesssim \|\widehat{\zeta} - \zeta^\star\|_\infty + \|\mathcal{J}'_{P^\star}(w^\star - w)\|_2
\end{aligned}$$

where $g_1 = ((\mathcal{J}_{P^\star} - \mathcal{J}_{\widehat{P}})'(w_{\widehat{P}} - w))/\|(\mathcal{J}_{P^\star} - \mathcal{J}_{\widehat{P}})'(w_{\widehat{P}} - w)\|_2$, $g_2 = (\mathcal{J}'_{P^\star}(w_{\widehat{P}} - w_{P^\star}))/\|\mathcal{J}'_{P^\star}(w_{\widehat{P}} - w_{P^\star})\|_2$. The last inequality uses (i) and (ii) with the fact that $\|g_1\|_\infty < \infty$ and $\|g_2\|_\infty < \infty$ as the w terms are bounded by our premise. Therefore, taking min over w and using Assumption I.1, we have $\varepsilon_A \lesssim \|\widehat{\zeta} - \zeta^\star\|_\infty + \varepsilon_{\mathsf{WReal}}$.

For approximate completeness (ε_B): for any $w \in \mathcal{W}$ and $f \in \mathcal{F}$, we have

$$\begin{aligned}
&\|f - \mathcal{J}'_{\widehat{P}}(w - w_{\widehat{P}})\|_2 \\
&\le \|f - \mathcal{J}'_{P^\star}(w - w_{P^\star})\|_2 + \|(\mathcal{J}_{P^\star} - \mathcal{J}_{\widehat{P}})'(w - w_{P^\star})\|_2 + \|\mathcal{J}'_{P^\star}(w_{\widehat{P}} - w_{P^\star})\|_2 \\
&\lesssim \|f - \mathcal{J}'_{P^\star}(w - w_{P^\star})\|_2 + \|\widehat{\zeta} - \zeta^\star\|_\infty,
\end{aligned}$$

for the same reason as ε_A as the error terms are the same. Thus, $\varepsilon_B \lesssim \|\widehat{\zeta} - \zeta^\star\|_\infty + \varepsilon_{\mathsf{WComp}}$.

In sum, we have shown that the misspecification is at most $\mathcal{O}(\|\widehat{\zeta} - \zeta^\star\|_\infty + \varepsilon_{\mathsf{WReal}} + \varepsilon_{\mathsf{WComp}})$. Therefore, [59, Theorem 6.1 and Appendix C] ensures that w.p. $1-\delta$, our learned $\widehat{w}$ satisfies,

$$\left\|\mathcal{J}'_{\widehat{P}}(\widehat{w} - w_{\widehat{P}})\right\|_2 \lesssim \varepsilon_n^{\mathcal{W}} + \|\widehat{\zeta} - \zeta^\star\|_\infty + \varepsilon_{\mathsf{WReal}} + \varepsilon_{\mathsf{WComp}} + \sqrt{\log(1/\delta)/n}.$$

Concluding the proof. The final step is to translate the above guarantee to $\|\mathcal{J}'_{P^\star}(\widehat{w} - w_{P^\star})\|_2$. The following shows that the switching cost is $\mathcal{O}(\|\widehat{\zeta} - \zeta^\star\|_\infty)$ as before:

$$\begin{aligned}
&\|\mathcal{J}'_{P^\star}(\widehat{w} - w_{P^\star})\|_2 \\
&\leq \|(\mathcal{J}_{P^\star} - \mathcal{J}_{\widehat{P}})'(\widehat{w} - w_{P^\star})\|_2 + \|\mathcal{J}'_{\widehat{P}}(\widehat{w} - w_{\widehat{P}})\|_2 + \|\mathcal{J}'_{\widehat{P}}(w_{\widehat{P}} - w_{P^\star})\|_2 \\
&\lesssim \varepsilon_n^{\mathcal{W}} + \|\widehat{\zeta} - \zeta^\star\|_\infty + \varepsilon_{\mathsf{WReal}} + \varepsilon_{\mathsf{WComp}} + \sqrt{\log(1/\delta)/n}.
\end{aligned}$$

This concludes the proof. □

Lemma I.3 (Visitation performance-difference)**.** *Let $P, U : \mathcal{S} \to \mathbb{R}_+$ be non-negative measures, which should be thought of as transitions in a discounted Markov chain. Assume U satisfies $\sum_{s'} U(s' \mid s) \leq 1$. Define $d_U = (1-\gamma)\sum_{h=1}^{\infty} \gamma^{h-1} d_U^h$, where $d_U^h = \int_{s_1, s_2, \dots, s_{h-1}} d_1(s_1) U(s_2 \mid s_1) \dots U(s \mid s_{h-1}) \mathrm{d}s_{1:h-1}$. Assume the same for P.*

Let $\mathcal{F} \subset \mathcal{S} \to \mathbb{R}$ be a function class that satisfies $f \in \mathcal{F} \implies g(s) = \mathbb{E}_{s' \sim P(s)}[f(s')] \in \mathcal{F}$, i.e., closed under projection with P. Then, define the integral (probability) metric $\|P - U\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |(\mathbb{E}_P - \mathbb{E}_U)[f(s)]|$. Then we have,

$$\|d_P - d_U\|_{\mathcal{F}} \leq \frac{\gamma}{1-\gamma} \mathbb{E}_{d_U} \|P(\cdot \mid s) - U(\cdot \mid s)\|_{\mathcal{F}}. \tag{18}$$

Proof. Recall Bellman's flow, which is $d_P(s) = (1-\gamma) d_1(s) + \gamma \mathbb{E}_{\widetilde{s} \sim d_P} P(s \mid \widetilde{s})$. Fix any $f \in \mathcal{F}$. The initial state distributions cancel, so we have,

$$\begin{aligned}
&|(\mathbb{E}_{d_P} - \mathbb{E}_{d_U})[f(s)]| \\
&= \left|\gamma \mathbb{E}_{\widetilde{s} \sim d_P} \mathbb{E}_{s \sim P(\cdot|\widetilde{s})}[f(s)] - \gamma \mathbb{E}_{\widetilde{s} \sim d_U} \mathbb{E}_{s \sim U(\cdot|\widetilde{s})}[f(s)]\right| \\
&\leq \left|\gamma \mathbb{E}_{\widetilde{s} \sim d_P} \mathbb{E}_{s \sim P(\cdot|\widetilde{s})}[f(s)] - \gamma \mathbb{E}_{\widetilde{s} \sim d_U} \mathbb{E}_{s \sim P(\cdot|\widetilde{s})}[f(s)]\right| \\
&+ \left|\gamma \mathbb{E}_{\widetilde{s} \sim d_U} \mathbb{E}_{s \sim P(\cdot|\widetilde{s})}[f(s)] - \gamma \mathbb{E}_{\widetilde{s} \sim d_U} \mathbb{E}_{s \sim U(\cdot|\widetilde{s})}[f(s)]\right| \\
&\leq \gamma \left|(\mathbb{E}_{\widetilde{s} \sim d_P} - \mathbb{E}_{\widetilde{s} \sim d_U})[\mathbb{E}_{s \sim P(\cdot|\widetilde{s})} f(s)]\right| + \gamma \mathbb{E}_{\widetilde{s} \sim d_U} \left|\left(\mathbb{E}_{s \sim P(\cdot|\widetilde{s})} - \mathbb{E}_{s \sim U(\cdot|\widetilde{s})}\right)[f(s)]\right|.
\end{aligned}$$

Thus, taking supremum over $\mathcal{F}$, we have

$$\begin{aligned}
&\|d_P - d_U\|_{\mathcal{F}} \\
&\leq \gamma \sup_{f \in \mathcal{F}} \left|(\mathbb{E}_{\widetilde{s} \sim d_P} - \mathbb{E}_{\widetilde{s} \sim d_U})[\mathbb{E}_{s \sim P(\widetilde{s})} f(s)]\right| + \gamma \mathbb{E}_{\widetilde{s} \sim d_U} \sup_{f \in \mathcal{F}} \left|\left(\mathbb{E}_{s \sim P(\cdot|\widetilde{s})} - \mathbb{E}_{s \sim U(\cdot|\widetilde{s})}\right)[f(s)]\right| \\
&= \gamma \|d_P - d_U\|_{\mathcal{F}} + \gamma \mathbb{E}_{\widetilde{s} \sim d_U} \|P(\cdot \mid \widetilde{s}) - U(\cdot \mid \widetilde{s})\|_{\mathcal{F}}. \qquad (\mathcal{F} \text{ closed under } P\text{-projection})
\end{aligned}$$

Rearranging terms finishes the proof. □

If $\mathcal{F}$ is the class of functions with $\|f\|_\infty \leq 1$, then this recovers the TV distance, which gives,

$$\|d_P - d_U\|_{\mathsf{TV}} \leq \frac{\gamma}{1-\gamma} \mathbb{E}_{d_U} \|P(\cdot \mid s) - U(\cdot \mid s)\|_{\mathsf{TV}}. \tag{19}$$

This generalizes Lemma E.3 of [1] to infinite horizon.

J Validity Guarantees for Orthogonal Estimator

Our orthogonal estimator has additional desirable properties such as *validity* when some nuisances are misspecified. Specifically, the bounds returned by our orthogonal estimator will be asymptotically valid, though possibly loose, when some nuisances are inconsistent, *i.e.*, do not converge to the their true values. Below, we detail conditions under which we achieve validity. To be concise, we focus on the $-$ case as the $+$ case is symmetric.

Validity with correct $Q^\pm$. If $\widehat{Q} = Q^\pm$, we obtain valid bounds even if w, β are inconsistent.

Lemma J.1. *For any w, β, we have $\mathbb{E}[\psi(s,a,s';Q^-,\beta,w)] \le V^-_{d_1}$ with equality when $\beta = \beta^-_\tau$.*

Validity with $Q = \mathcal{T}^\pm_\beta Q$. Even if $\widehat{Q}$ is misspecified, we still have a valid bound if it solves a Bellman-type equation of the dual CVaR form. For a $\beta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define:

$$\begin{aligned}\mathcal{T}^\pm_\beta f(s,a) := r(s,a) + \gamma\Lambda^{-1}(s,a)\mathbb{E}[f(s',\pi_{\mathrm{t}}) \mid s,a] \\ + \gamma(1-\Lambda^{-1}(s,a))\mathbb{E}\big[\beta(s,a) + \tau^{-1}(s,a)(f(s',\pi_{\mathrm{t}}) - \beta(s,a))_\pm \mid s,a\big].\end{aligned}$$

Lemma J.2. *Fix any w, β. If $Q^\pm_\beta = \mathcal{T}^\pm_\beta Q^\pm_\beta$, then $\mathbb{E}[\psi(s,a,s';Q^-_\beta,\beta,w)] \le V^-_{d_1}$.*

Remark J.3. Lemmas J.1 and J.2 are dual to each other: in Lemma J.1, the plug-in is consistent while the debiasing correction errs in the valid direction (*i.e.*, ≥ 0 for $+$ and ≤ 0 for $-$). In Lemma J.2, the plug-in is valid while the debiasing correction has expectation zero.

J.1 Proofs for validity

Lemma J.1. *For any w, β, we have $\mathbb{E}[\psi(s,a,s';Q^-,\beta,w)] \le V^-_{d_1}$ with equality when $\beta = \beta^-_\tau$.*

Proof.

$$\begin{aligned}\mathbb{E}[\psi(s,a,s';Q^-,\beta,w)] &\le (1-\gamma)\mathbb{E}_{d_1}[V^-_\beta(s_1)] + \mathbb{E}[w(s,a)\big(Q^-(s,a) - \mathcal{T}^-_{\mathrm{CVaR}}Q^-(s,a)\big)] \\ &= V^-_{d_1} + 0 = V^-_{d_1},\end{aligned}$$

where the inequality comes from the fact that β is sub-optimal for $\mathbb{E}[\beta(s,a) + \tau^{-1}(V^-(s') - \beta(s,a))_-]$. The same proof applies for Q^+. □

We now prove Lemma J.2. First, we show that the $\mathcal{T}_\beta$ perspective gives rise to a dual definition of $Q^\pm$ (dual to Eq. (2)).

Lemma J.4.

$$Q^+(s,a) = \arg\min\nolimits_{\beta:Q_\beta=\mathcal{T}^+_\beta Q_\beta} Q_\beta(s,a), \quad Q^-(s,a) = \arg\max\nolimits_{\beta:Q_\beta=\mathcal{T}^-_\beta Q_\beta} Q_\beta(s,a).$$

Proof. Unroll $Q^-(s,a) = r(s,a) + \gamma \inf_{U\in\mathcal{U}(P)} \mathbb{E}_U[r(s',a') + \gamma \inf_{U\in\mathcal{U}(P)} \mathbb{E}_U[\dots]]$, replacing each $\inf_{U\in\mathcal{U}(P)}$ with the convex combination of $\mathbb{E}$ and CVaR from Lemma 3.1. Then, write each CVaR using the dual form, *i.e.*, $\max_\beta\{\beta(s,a) + \tau^{-1}(s,a)\mathbb{E}[(\cdots - \beta(s,a))_+]\}$. By s,a-rectangularity, the scalar $\max_\beta$ separates per s,a, so we can pull all the maxes out front as a max over $\beta(s,a)$ functions. Note that not all $\beta(s,a)$ functions have a well-defined infinite sum in this manner, as $\mathcal{T}_\beta$ is not always a contraction. The condition $Q_\beta = \mathcal{T}^-_\beta Q_\beta$ exactly characterizes when this unrolling is well-defined. Thus, Q^- is exactly the minimum Q_β whenever this procedure of unrolling with β is well-defined. This concludes the proof. □

Lemma J.2. *Fix any w, β. If $Q^\pm_\beta = \mathcal{T}^\pm_\beta Q^\pm_\beta$, then $\mathbb{E}[\psi(s,a,s';Q^-_\beta,\beta,w)] \le V^-_{d_1}$.*

Proof.

$$\mathbb{E}[\psi(s,a,s';Q^-_\beta,\beta,w)] = (1-\gamma)\mathbb{E}_{d_1}[V^-_\beta(s_1)] + 0 \le V^-_{d_1}.$$

The first equality is because the correction term is $\mathcal{T}^-_\beta Q^-_\beta - Q^-_\beta$, which is zero since Q^-_β is a fixed point. The inequality is due to Lemma J.4. □

K Additional Experiment Details

K.1 Environment

We consider a simple MDP with a one-dimensional state space $\mathcal{S} = [0, 5]$, a binary action space $\mathcal{A} = \{0, 1\}$, reward function

$$r(s, a) = \frac{26 - s^2 - \mathbb{I}[a = 1]}{26},$$

which we note takes values in the range $[0, 1]$, and with transitions given by

$$\begin{aligned} P(\cdot \mid s, a = 0) &= \text{UnifClip}[s - 0.2,\ s + 1] \\ P(\cdot \mid s, a = 1) &= \text{UnifClip}[0.2s - 0.02,\ s + 0.5], \end{aligned}$$

where UnifClip$[a, b]$ denotes a uniform distribution between $\max(a, 0)$ and $\min(b, 5)$. In addition, the environment always starts in initial state $s_0 = 2$. Essentially, this is a simple control environment, where high rewards are obtained by maintaining state as close to zero as possible, the action $a = 1$ is a control action that (in expectation) moves the state closer to zero, and which occurs a small reward cost, and the action $a = 0$ is a passive action that allows the state to freely drift (with an overall drift away from zero).

K.2 Target Policy

We focus on estimating the worst-case policy value $V^-_{d_1}$ for the simple threshold-based target policy π_t which takes action $a = 1$ when $s \geq 2$, and $a = 0$ whenever $s < 2$.

K.3 Logging Policy and Data Sampling Procedure

We sample data using an evaluation policy π_b which is an ϵ-smoothed threshold policy similar to π_t. Specifically, π_b takes action $a = 1$ when $s \geq 1.5$ with probability 0.95, and takes action $a = 0$ when $s < 1.5$ with probability 0.95. We obtain a dataset $\{s_i, a_i, s'_i, r_i\}$ by first rolling out with π_b for 1000 burn-in time steps, and then sampling the tuple (s, a, s', r) every 10 time steps. For each replication of our experiment, we sample 10,000 tuples in total.

K.4 Calculation of True Worst-case Policy Values

A major challenge in studying robust policy value estimation is that, even with ground truth knowledge of the MDP and/or access to a simulator, it may be intractable to estimate the robust policy values $V^{\pm}_{d_1}$. Fortunately, the above environment has the desirable property that we can analytically compute the best/worst-case transition distributions allowed by our sensitivity model, since no matter what policy π_t the agent is acting with, it always strictly prefers transitions to smaller states. In detail, suppose that for some state, action pair (s, a) we have $P(\cdot \mid s, a) = \text{Unif}[x, y]$, for some $0 \leq x \leq y \leq 5]$. Then, letting $\alpha = 1/(1 + \Lambda(s, a))$, it is easy to verify that the worst case transition kernel is given by

$$U^-(\cdot \mid s, a) = (1 - \Lambda^{-1}(s, a))\text{Unif}[y - \alpha(y - x), y] + \Lambda^{-1}(s, a)\text{Unif}[x, y].$$

That is, the worst case transition kernel is given by a mixture of two uniform distributions. Therefore, we can easily simulate rollouts with the best/worst case transition kernels, and accurately estimate the robust policy values. This allows us to validate our methodology in this synthetic environment. Specifically, for each $\Lambda(s, a)$ we experiment with, we can compute the corresponding ground truth $V^-_{d_1}$ up to arbitrary precision via Monte Carlo sampling, by rolling out trajectories with π_t in the adversarial MDP according to the above worst-case transition kernel.

Note as well that if one wanted to estimate the best-case policy value, analogous reasoning would give us

$$U^+(\cdot \mid s, a) = (1 - \Lambda^{-1}(s, a))\text{Unif}[x, x + \alpha(y - x)] + \Lambda^{-1}(s, a)\text{Unif}[x, y].$$

However, in our experiments we only concern ourselves with worst-case policy value estimation.

K.5 Nuisance Estimation

We instantiate slight variations of Algorithms 1 and 2 using neural nets for the classes $\mathcal{Q}$, $\mathcal{B}$, and $\mathcal{W}$ used for fitting Q^-, β^-, and w^- respectively, and linear sieves for the corresponding critic class $\mathcal{Q}$ that we perform maximization over for the minimax estimation of w^-. Specifically, we grow the linear sieve for the critic class in a data-driven way, as follows: at each step k of the respective algorithm, we compute the best response $q_k \in \mathcal{Q}$ to the previous iterate solution $w_k \in \mathcal{W}$ by optimizing over a neural net class, and then we append this best-response function to the set of functions in our linear sieve for the corresponding critic class. Full exact nuisance estimation details necessary for reproducibility will be available in our code release.

K.6 Estimators

We estimate the worst-case policy value using three different estimators:

- **Q**: Direct estimator given by:
$$\widehat{V}^-_{d_1} = \widehat{Q}^-(s_1, \pi_t(s_1))\,,$$
where s_1 is the deterministic initial state.
- **W**: Importance sampling-style estimator using $\hat{w}^-$, which is given by:
$$\widehat{V}^-_{d_1} = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}^-(s_i, a_i)\widehat{\xi}_i r_i\,,$$
where
$$\widehat{\xi}_i = \Lambda^{-1} + (1-\Lambda^{-1})(1+\Lambda)\mathbb{I}\left[\widehat{V}^-(s'_i) \leq \widehat{\beta}^-(s_i, a_i)\right]\,.$$
- **Orth**: Our orthogonal estimator using EIF, given by
$$\widehat{V}^-_{d_1} = \frac{1}{n}\sum_{i=1}^{n} \psi(s_i, a_i, s'_i; \widehat{Q}^-, \widehat{\beta}^-, \widehat{w}^-)\,.$$

Note as well that we used a simpler data splitting procedure rather than the cross-fitting procedure described in Algorithm 3. Specificallly, we used the first 10,000 tuples for estimating nuisances, and the second 10,000 tuples for the final estimators. This was done for the sake of computational ease in running experiments with many replications, and was performed in the same way for all methods.

In addition, for extra robustness, in each experiment replication we ran the nuisance estimation pipeline 5 times (on the same fixed sampled dataset), and took the 80th percentile policy value estimates, since the estimators tend to under-estimate the true policy value by design, with greater under-estimation when the nuisance estimates are less well optimized.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] .

   Justification: Yes, we provide complete proofs for our theorems and describe detailed empirical validation for our proposed algorithms.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: Yes, we discussed where our assumptions may fail and settings not captured by the current paper, which we believe are directions for future research.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Yes, we provide full assumptions in the main paper and the complete proofs are written in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes] .

   Justification: Yes, our experimental section includes all details needed to reproduce the main experimental results. We will also open-source our code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes] .

Justification: Yes, we will open-source code to reproduce our experiment results on GitHub once our organization's internal code-review process is approved.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Yes, please see our experimental section for all training and evaluation details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes] .

Justification: Yes, our experiments are replicated over multiple seeds and we report the confidence intervals.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes] .

   Justification: Yes, this paper is mostly focused on theory and our experiment is a proof of concept and can be run on a standard GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes] .

   Justification: Yes, we have reviewed the code of ethics and believe our research conforms to it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA] .

    Justification: This paper is about foundational research not tied to particular applications so we do not feel the need to highlight any societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: This paper is about foundational research not tied to particular applications so we do not feel the need to highlight any risks for misuse here.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA] .

    Justification: The paper does not use any existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.