Structured Difference-of-Q via Orthogonal Learning

Defu Cao
University of Southern California

Angela Zhou
University of Southern California

Abstract

Offline reinforcement learning is important in many settings with available observational data but the inability to deploy new policies online due to safety, cost, and other concerns. Many recent advances in causal inference and machine learning target estimation of "causal contrast" functions such as CATE, which is sufficient for optimizing decisions and can adapt to potentially smoother structure. We develop a dynamic generalization of the R-learner (Nie et al., 2021; Lewis and Syrgkanis, 2021) for estimating and optimizing the difference of Q^π-functions, $Q^\pi(s,a) - Q^\pi(s,a_0)$, for potential discrete-valued actions a, a_0, which can be used to optimize multiple-valued actions without loss of generality. We leverage orthogonal estimation to improve convergence rates, even if Q and behavior policy (so-called nuisance functions) converge at slower rates and prove consistency of policy optimization under a margin condition. The method can leverage black-box estimators of the Q-function and behavior policy to target estimation of a more structured Q-function contrast, and comprises of simple squared-loss minimization.

1 INTRODUCTION

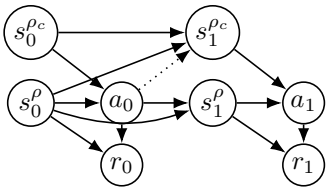
Sequential decision-making is a cornerstone of many real-world applications, from personalized healthcare (Jayaraman et al., 2024) to autonomous systems (Levine et al., 2020), where optimizing dynamic treatment regimes or action policies is critical. However, in many domains, only observational data is available, limiting the ability to conduct online exploration (Gupta et al., 2023). Offline reinforcement learning (RL) enables evaluating and optimizing sequential decision rules in such settings, where the goal is to learn optimal policies from historical data without the ability to actively interact with the environment.

Offline reinforcement learning shares deep connections with causal inference. Offline RL connects closely to causal inference, where recent work emphasizes learning causal contrast functions such as the conditional average treatment effect (CATE) (Wager and Athey, 2018; Foster and Syrgkanis, 2019; Künzel et al., 2019; Kennedy, 2020). The causal contrast (i.e. the difference that actions make on outcomes) may be smoother or sparser than the main effects (what happens under either action by itself, Q^π). The causal contrast also admits opportunities for more robust and efficient estimation leveraging orthogonal and double machine learning (Chernozhukov et al., 2018; Kennedy, 2022).

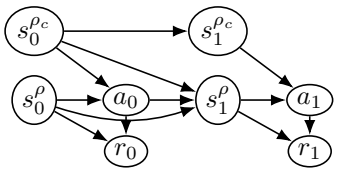
In this work, building on recent advances in heterogeneous treatment effect estimation, we focus on estimating analogous causal contrasts for offline reinforcement learning, namely $\tau_t^\pi(s; a, a_0) = Q_t^\pi(s,a) - Q_t^\pi(s,a_0)$, for possible actions a, a_0 in the action space $\mathcal{A}$. While we initially focus on binary actions for clarity, our framework naturally extends to multiple actions. This is closely related to, but different from advantage functions in reinforcement learning, defined as $Q^\pi(s,a) - V^\pi(s)$, the advantage of taking action a beyond the policy. For brevity, we denote the binary-action difference-of-Q, $\tau_t^\pi(s) = Q_t^\pi(s,1) - Q_t^\pi(s,0)$.

The sequential setting offers even more motivation to target estimation of the contrast: additional structure can arise from sparsity patterns induced by the joint (in)dependence of rewards and transition dynamics on the (decompositions of) the state variable. Recent works point out this additional structure (Wang, Xiao, Zhu, and Stone, Wang et al.; Wang et al., 2022). For example a certain transition-reward factorization, first studied by (Dietterich et al., 2018), admits a sparse Q-function contrast (Pan and Schölkopf, 2023). (Zhou, 2024b) proposes a variant of the underlying blockwise pattern that also admits sparse optimal Q functions and policies. Section 1 illustrates how both of these structures have very different conditional independence

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).



(a) Reward-relevant/irrelevant factorization (Zhou, 2024b). Dotted $a_t \to s_{t+1}^{\rho_c}$ is optional. The optimal Q is sparse in s^{ρ} (i.e. doesn't change if $s_0^{\rho_c}$ changes).



(b) Exogenous–Endogenous MDP (Dietterich et al., 2018). Q not sparse in s^{ρ}, but $Q(s,1) - Q(s,0)$ is.

Table 1: Comparison of Desiderata and Methods.

Features	DROPE[1]	Dyn-R[2]	DAE[3]	FQE[4]	Diff-Q (Ours)
Difference of Q function	✗	✗	✓	≈	✓
Orthogonal estimation	✓	✓	✗	✗	✓
Avoids Multiplied IS	≈	✓	✓	✓	✓
Convex loss	n/a	✓	✗	✓	✓

assumptions. Methods designed assuming one model is correct may not perform well if it is not. However, both structures imply that the difference-of-Q functions is sparse in an "endogenous" state component. This illustrates our broader motivation: directly estimating Q-contrasts adapts to underlying structure, such as sparsity or smoothness, even when the individual Q-functions are more complex.

Beyond our broad motivation, our approach has some appealing methodological aspects of independent interest: 1) evaluating the estimator doesn't require plugging in unstable *inverse* propensity weights, 2) it is based on a loss-minimization framework, and 3) targeting the entire difference-in-Q function helps avoid challenges of policy optimization with policy-dependent nuisance functions (Zhang et al., 2013; Xu et al., 2021). In this work, we extend the R-learner approach (Nie and Wager, 2021; Lewis and Syrgkanis, 2021) to sequential settings. We estimate and optimize Q-function contrasts, bridging recent advances in causal inference with offline reinforcement learning. Our main result is that under weaker conditions, that the Q-functions and estimation of behavior policies are $o_p(n^{-\frac{1}{4}})$ convergent in root-mean-squared error, and standard structural assumptions of Bellman-complete Q, well-specified difference-of-Q, and concentrability, we obtain worst-case $O_p(n^{-\frac{1}{2}})$ rates of convergence for estimating the difference of Q functions and attaining the optimal policy value. The loss minimization framework is flexible, which enables extensions like mutual-information-based regularization to adapt to simpler structure in the difference-of-Q function, and the benefits of adapting to structure in synthetic examples.

2 RELATED WORK

There is a large literature on offline reinforcement learning (evaluation and optimization) (Jin et al., 2021; Xie et al., 2021), including approaches that leverage importance sampling or introduce marginalized versions (Jiang and Li, 2016; Thomas et al., 2015; Kallus and Uehara, 2019a; Liu et al., 2018). For Markov decision processes, other papers study statistically semiparametrically efficient or doubly-robust estimation, but of the *averaged policy value* $\mathbb{E}[V_1^{\pi^e}(S_1)]$, rather than MSE convergence of the difference-of-Q function as we do here (Kallus and Uehara, 2019a,b; Xie et al., 2023). The literature on dynamic treatment regimes (DTRs) studies a method called advantage learning (Schulte et al., 2014), although DTRs in general lack reward at every timestep, whereas we are particularly motivated by sparsity implications that arise jointly from reward and transition structure. Beyond policy value estimation, we seek the entire contrast function.

We discuss crucial distinctions from other works studying advantage functions. Advantage functions appear in RL and dynamic treatment regimes (Neumann and Peters, 2008; Murphy, 2003). Unlike advantage functions in RL and DTRs, our difference-of-Q is π_t policy-independent at time t, unlike advantage functions, simplifying optimization. Pan and Schölkopf (2023, 2024) propose direct advantage estimation that avoids Q-functions but requires a challenging nonconvex constraint. Farias et al. (2022) develop an estimator called Differences-In-Q for the difference in average value under all-treat or all-control. The estimator averages a difference of Q functions for variance reduction. Our method can be used, although they target a different average policy value estimand. Closest to our goal is Shi et al. (2022), who derive a pseudo-outcome for estimating Q-contrasts in the infinite-horizon setting. Our method differs in focusing on the finite-horizon case, leveraging an orthogonal R-learner formulation that provides improved statistical guarantees and practical convex estimation.

At a high level, our method is similar to the dynamic DR learner studied in (Lewis and Syrgkanis, 2021) in that we extend the R-learner identification approach to a sequential setting, although the estimand is quite different. They generalize structural nested-mean models by estimating "blip-to-zero" functions but only consider heterogeneity based on a fixed *initial* state and dy-

namic treatment regimes with terminal rewards. Note that the (single-stage) R-learner loss function is an overlap-weighted (Li et al., 2019) regression against the doubly-robust score (DR-learner (Kennedy, 2020)). (See (Morzywolek et al., 2023; Chernozhukov et al., 2024)). We do make a margin assumption to relate convergence of Q-function contrasts to policy value, analogous to (Shi et al., 2022; Hu et al., 2024).

3 METHOD

Problem Setup: We consider a finite-horizon Markov Decision Process, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, P, \gamma, T)$ of state space $\mathcal{S}$, discrete action space $\mathcal{A}$, reward function $r : \mathcal{S} \times \mathcal{A} \to \mathcal{R}$, transition probability $P\colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ is the set of distributions over $(\mathcal{S})$, discount factor $0 \leq \gamma < 1$, and time horizon of T steps. We let $t = 1, \dots, T$ index timesteps. We let the state spaces $\mathcal{S} \subseteq \mathbb{R}^d$ be continuous, and assume the action space $\mathcal{A}$ is finite. Following causal conventions we denote $\pi(a \mid s)$ as the probability of taking action a in state s; at times we omit dependence on function arguments referring to the policy function π. Capital letters denote random variables $(S_t, A_t, \dots)$, lower case letters s, a denote evaluation at a generic value.

The value function is $V_t^\pi(s) := \mathbb{E}_\pi[\sum_{t'=t}^T \gamma^{t'-t} R_{t'} \mid S_t = s]$ where $\mathbb{E}_\pi$ denotes expectation under the joint distribution induced by the MDP $\mathcal{M}$ running policy π. The $Q-$function is the (s, a)-conditional expectation of discounted future rewards. The optimal value and q-functions are denoted V^*, Q^* under the optimal policy. We assume we are working in a Markov decision process, so that the joint distribution of the data-generating process of the observed data factorizes as follows:

$$\begin{aligned} &p(S_1, A_1, R_1, \dots, S_T) \\ &= p_{S_1}(S_1) \textstyle\prod_{t=1}^T \pi_t^b(A_t \mid S_t) p\left(S_{t+1}, R_t \mid S_t, A_t\right) \end{aligned} \quad (1)$$

Under the behavior policy π_b, actions A_t were taken with probability depending on observed states S_t alone (and not on any unobserved states). It corresponds to assuming MDP rather than POMDP structure, and is a common implicit assumption in RL.[5]

At times we omit the function evaluation where it is clear from context, for example referring to the reward function $R_t(S_t, A_t)$ as R_t in longer expressions. We focus on estimating the difference of Q-functions (each under the same policy), $\tau_t^\pi(s) = Q_t^\pi(s, 1) - Q_t^\pi(s, 0)$. We focus on the offline reinforcement learning setting with a historical dataset of n offline trajectories, $\mathcal{D} = \{(S_t^i, A_t^i, R_t^i, S_{t+1}^i)_{t=1}^T\}_{i=1}^n$, where actions were sampled under a behavior policy π^b. Notationally: following convention in statistical papers on causal inference, we denote the $\mathcal{L}_2(P)$-norm $\|f(X)\|_2 := \mathbb{E}[|f(X)|^2]^{1/2}$; expectations and norms are under the observational behavior distribution unless otherwise indicated.

Policy Evaluation (Identification): First we overview deriving our method. The arguments are broadly a generalization of the so-called residualized R-learner (Nie and Wager, 2021); (Lewis and Syrgkanis, 2021) considers a similar generalization for structural nested mean models without state-dependent heterogeneity. For this section, we discuss the true population Q, τ, etc., functions without decoration with $\circ$, introduced later in the context of estimation.

Next we discuss identification and our estimator. For brevity, we denote Q-functions under some policy π from time $t+1$ onwards as Q_t^π. Our goal is to estimate the difference of Q functions:

$$\tau_t^\pi(S_t) := Q_t^\pi(S_t, 1) - Q_t^\pi(S_t, 0). \quad (2)$$

To begin, note that the Q-function satisfies: $Q_t^\pi(S_t, A_t) = \mathbb{E}[R_t + \gamma Q_{t+1}^\pi(S_{t+1}, A_{t+1}) \mid S_t, A_t]$.

By Markovianity, eqn. (1) and by definition of $\tau_t^\pi(S_t)$, we obtain that:

$$\begin{aligned} &\mathbb{E}[R_t + \gamma Q_{t+1}^\pi(S_{t+1}, A_{t+1}) \\ &\quad - \{Q_t^\pi(S_t, 0) + A_t \tau_t^\pi(S_t)\} \mid S_t, A_t] = 0 \end{aligned}$$

Eqn. 1 grants that the above holds conditionally on (S_t, A_t) rather than the entire history. Analogous to the marginal outcome, denote the state-conditional value function under the behavior policy: $m^{\circ,\pi}(S_t) := V_t^{\pi_t^b, \pi_{t+1}}(S_t)$. By Markovianity, marginalizing the above equation over actions:

$$\begin{aligned} m_t^\pi(S_t) &:= \mathbb{E}_{\pi^b}[R_t + \gamma Q_{t+1}^\pi(S_{t+1}, A_{t+1}) \mid S_t] \\ &= Q_t^\pi(S_t, 0) + \pi_1^b(S_t) \tau_t^\pi(S_t) \end{aligned} \quad (3)$$

Differencing these, we obtain the identifying moment condition, satisfied by the true difference-of-Q function $\tau_t^\pi(S_t)$:

$$\begin{aligned} &\mathbb{E}[\{R_t + \gamma Q_{t+1}^\pi(S_{t+1}, A_{t+1}) - m_t^\pi(S_t)\} \\ &\quad - \{A - \pi_1^b(S_t)\} \tau_t^\pi(S_t) \mid S_t, A_t] = 0 \end{aligned} \quad (4)$$

The loss function. This identifying moment condition motivates our approach based on (potentially penalized) empirical risk minimization. We minimize the following loss function for τ over a regression function class $\mathcal{G}$. The loss function depends on Q and behavior policy π_b functions. Since they are not the final targets of analysis, the causal ML literature calls them

[5] We include it in the problem setup as it is an RL/MDP-native counterpart of sequential ignorability in causal inference, but without additional unnecessary or confusing potential outcomes notation.

Algorithm 1 Dynamic Residualized Difference-of-Q Optimization

Given: Partition of $\mathcal{D}$ into 3 folds, $\{\mathcal{D}_k\}_{k=1}^3$.
Estimate $\hat{\pi}_t^b$ on $\mathcal{D}_1$.
At time T: Set $\hat{Q}_T(s,a)=0$. Estimate m_T on $\mathcal{D}_1$ and $\hat{\tau}_T$ on $\mathcal{D}_{k(T)}$, where $k(t)=2$ if t is odd and $k(t)=3$ if t is even. Optimize, for two actions $\hat{\pi}_T(s)=\mathbb{I}[\hat{\tau}_T(s)>0]$.
for timestep $t=T-1,\dots,1$ **do**
  Estimate $Q^{\hat{\pi}_{t+1}}$ on $\mathcal{D}_1$. Evaluate $m_t^{\hat{\pi}_{t+1}}$. Estimate $\hat{\tau}_t^{\hat{\pi}_{t+1}}$ on $\mathcal{D}_{k(t)}$ by minimizing:

$$\hat{\tau}_t(\cdot)\in \text{argmin}_\tau \sum_{i\in\mathcal{D}_{k(t)}}\Big(R_t^i+\gamma\mathbb{E}_{a'\sim\hat{\pi}_{t+1}}[\hat{Q}_{t+1}^{\hat{\pi},(1)}\left(S_{t+1}^i,a'\right)\mid S_{t+1}]-\hat{m}_t^{\hat{\pi},(1)}(S_t^i)-\{A_t^i-\hat{\pi}_t^{b,(1)}(S_t^i)\}\tau_t(S_t^i)\Big)^2$$

  Optimal policy is greedy with respect to the difference-of-Q function. For two actions, $\hat{\pi}_t(s)=\mathbb{I}[\hat{\tau}_t^{\hat{\pi}_{t+1}}(s)>0]$. Else for multiple actions, $\hat{\pi}_t(s)\in\arg\max_{a'\in\mathcal{A}\setminus a_0}\hat{\tau}_t^{\hat{\pi}_{t+1}}(s,a')$ if $\max_{a'\in\mathcal{A}\setminus a_0}\hat{\tau}_t^{\hat{\pi}_{t+1}}(s,a')>0$, else a_0.
end for

"nuisance functions". Notationally, they are denoted as the nuisance vector $\eta=[\{Q_t^\pi\}_{t=1}^T,\{m_t^\pi\}_{t=1}^T,\{\pi_t^b\}_{t=1}^T]$.

$$\tau_t(\cdot)\in\underset{\tau\in\mathcal{G}}{\text{argmin}}\,\mathcal{L}(\tau,\eta),$$
$$\mathcal{L}_t^\pi(\tau,\eta):=\mathbb{E}\left[\left(\{R_t+\gamma\mathbb{E}_{a'\sim\pi}[Q_{t+1}^\pi(S_{t+1},a')\mid S_{t+1}]-m_t^\pi(S_t)\}-\{A-\pi_t^b(1\mid S_t)\}\cdot\tau(S_t)\right)^2\right] \tag{5}$$

Extension to multiple actions. So far we presented the method with $\mathcal{A}\in\{0,1\}$ for simplicity, but all methods in this paper will extend to the multi-action case. We give the approach discussed in (Nie and Wager, 2021). For multiple actions, fix an arbitrary baseline choice $a_0\in\mathcal{A}$, and for $a\in\mathcal{A}\setminus a_0$, define $\tau_t^\pi(s,a):=\tau_{a-a_0,t}^\pi(s)=Q_t^\pi(s,a)-Q_t^\pi(s,a_0)$. For $k\in|\mathcal{A}|$, let $\pi^b(k\mid S_t)=P(A_t=k\mid S_t)$. Then the analogous loss function is, where $\langle\cdot,\cdot\rangle$ is the dot product so that $\langle\vec{A}-\vec{\pi}^b(S_t),\tau_t^\pi(S_t)\rangle=\sum_{a\in\mathcal{A}\setminus a_0}\{(\mathbb{I}[A_t=a]-\pi^b(a\mid S_t))\tau_{a,t}^\pi(S_t)\}$:

$$\vec{\tau}_t(\cdot)\in\text{argmin}_\tau\{\mathbb{E}[(\{R_t+\gamma\mathbb{E}_{a'\sim\pi}[Q_{t+1}^\pi(S_{t+1},a')\mid S_{t+1}]-m_t^\pi(S_t)\}-\langle\vec{A}-\vec{\pi}^b(S_t),\vec{\tau}_{a_0,t}^\pi(S_t)\rangle)^2]\}.$$

Again, so far we have discussed identification assuming the true Q,m,π^b functions, etc. Next we discuss feasible estimation, and outside of this section we refer to the population-level true nuisance functions as $Q^{\pi,\circ},m^{\pi,\circ},\pi^{b,\circ},\tau^{\pi,\circ}$.

Feasible Estimation: We optimize a feasible version of the sequential loss minimization in eqn. (5) with estimated nuisances (see Algorithm 3). Given an evaluation policy π^e, first fit the $\hat{Q}^{\pi^e}$ function and the behavior policy. Evaluate the value function via integrating/summing $\hat{Q}^{\pi^e}$ over the empirical distribution of actions, $a\sim\pi^b$, so that $\hat{m}(s)=\mathbb{E}_{\pi_t^b}[R_t+\gamma\hat{Q}_{t+1}^{\pi^e}(S_{t+1},A_{t+1})\mid S_t=s]$. Then, evaluate the loss function in eqn. (5) and estimate τ_t. The Q function can be estimated with standard approaches such as fitted-Q-evaluation (Le et al., 2019; Chakraborty and Murphy, 2014; Duan et al., 2021), conditional moment restrictions/GMM (Kallus and Uehara, 2019a). Estimating the behavior policy can be done with standard probabilistic classification, although it is often known.

Cross-fitting and Policy Optimization: We also introduce cross-fitting for policy evaluation and optimization. We split the dataset $\mathcal{D}$ into K folds (preserving trajectories, i.e. randomizing over trajectory index i), and learn the nuisance function η^{-k} on $\{\mathcal{D}_{k'}\}_{k'\in\{[K]\setminus k\}}$. (Sometimes for clarity we denote the nuisance function $\eta^{(-k)}$ instead.) In the loss-function, we evaluate the nuisance function η^{-k} on data in the held-out kth fold.

Estimating the difference-of-Q function. Putting the above together, we briefly state our procedure for estimating the difference-of-Q function (see appendix for full algorithm box). The inputs are the evaluation policy π^e, and for sample splitting, data partitions of $\mathcal{D}$ into K folds, $\{\mathcal{D}_k\}_{k=1}^K$. Let $\mathbb{E}_{a'\sim\pi^e}[Q_{t+1}^{\pi^e}(S_{t+1},a')\mid S_{t+1}]=\sum_{a'}\pi^e\left(a'\mid S_{t+1}^i\right)\hat{Q}_{t+1}^{\pi^e}\left(S_{t+1}^i,a'\right)$. Then:

- On $\mathcal{D}_k$, estimate $\hat{Q}^{\pi^e,k}$, behavior policy $\hat{\pi}_t^{b,k}$, therefore $\hat{m}^{\pi^e,k}$.
- For timestep $t=T,\dots,1$, optimize $\hat{\tau}_t\in\text{argmin}_\tau$

$$\left\{\sum_{k=1}^K\sum_{i\in\mathcal{D}_k}\Big(R_t^i+\gamma\mathbb{E}_{a'\sim\pi^e}[Q_{t+1}^{\pi^e}(S_{t+1},a')\mid S_{t+1}]-\hat{m}_t^{\hat{\pi},-k}(S_t^i)-\{A_t^i-\hat{\pi}_t^{b,-k}(S_t^i)\}\tau_t(S_t^i)\Big)^2\right\}$$

Policy optimization. The sequential loss minimization approach also admits an policy optimization procedure. The policy is greedy with respect to the estimated τ_t. We describe the algorithm in Algorithm 1. We use a slightly different cross-fitting approach for policy optimization. We introduce an additional fold, upon which

we alternate estimation of $\hat{\tau}_t$. So, overall we use three folds: one for estimating nuisance functions η, and the other two for estimating $\hat{\tau}_t^{\hat{\underline{\pi}}_{t+1}}$. On these two other folds, between every timestep, we alternate estimation of $\hat{\tau}_t$ on one of them, in order to break dependence between the estimated optimal forwards policy $\hat{\underline{\pi}}_{t+1}$ and $\hat{\tau}_t$ (and therefore the greedy policy $\hat{\pi}_t$).

4 ANALYSIS

We study the improved statistical rates of convergence from orthogonal estimation for policy evaluation (Theorem 4.1) and show that this implies convergent policy optimization (Theorem 4.2). Theorem 4.1 applies orthogonal statistical learning to our new estimand, for which we establish Neyman-orthogonality. Policy optimization is more challenging; the novelty of Theorem 4.2 is that estimation error from *policy-dependent nuisance functions* is of higher-order than the evaluation rates. Now that we discuss estimation rather than identification, we denote the true population functions with a $\circ$ superscript, i.e. $\tau_t^{\pi,\circ}$. Our analysis proceeds under the following assumptions.

Assumption 1 (Independent and identically distributed trajectories)**.** We assume that the data was collected under a stationary behavior policy, i.e. not adaptively collected from a policy learning over time.

Assumption 2 (Boundedness)**.** $V_t(s) \leq B_V, \tau(s) \leq B_\tau, \forall t, s$

Assumption 3 (Sup-norm concentrability)**.** Denote the marginal state-action distribution under a policy π as $d^\pi(s,a)$. There exists a constant C_∞ such that for any policy π (including non-stationary policies): $\forall \pi, s, a : \frac{d_\pi(s,a)}{d_{\pi^b}(s,a)} \leq C_\infty$.

Assumption 1 can be generalized with standard tools. Assumption 3, sup-norm concentrability, translates error bounds under one state distribution to another and includes a wide class of MDPs (Munos and Szepesvári, 2008). Recent works introduce weaker concentrability, for example with pessimism, we leave this for future work and focus on the key novelty of our work.

A key assumption for orthogonal estimation is that the product of estimation errors for the nuisance-functions converges quickly enough.

Assumption 4 (Product error rates on nuisance function evaluation)**.** Fix an evaluation policy π^e. Suppose the propensities and Q^{π^e} functions are $o_p(n^{-\frac{1}{4}})$ RMSE-consistent, i.e. $\mathbb{E}[\|\hat{\pi}_t^b - \pi_t^{b,\circ}\|_2] = o_p(n^{-\frac{1}{4}})$, and $\mathbb{E}[\|\hat{Q}_{t+1}^{\pi^e} - Q_{t+1}^{\pi^e,\circ}\|_2] = o_p(n^{-\frac{1}{4}})$.

We assume well-specification to simplify theorem statements, with general versions in the appendix.

Assumption 5 (Well-specification of τ)**.** $\min_{\tau \in \mathcal{G}} \mathcal{L}(\tau, \eta^0) = 0, \forall t$

Assumption 6 (Bellman completeness for Q^π)**.** There exists $\epsilon > 0$ such that, for all $t \in [T]$, where $\mathcal{T}^* f_{t+1}(s,a) = \mathbb{E}[r_t + \max_a f_{t+1}(S_{t+1}, a) \mid s, a]$, $\sup_{f_{t+1} \in \mathcal{F}_{t+1}} \inf_{f_t \in \mathcal{F}_t} \|f_t - \mathcal{T}^* f_{t+1}\|_2^2 \leq \epsilon$.

Assumption 5 posits the function class for estimating τ is well-specified. Meanwhile, Assumption 4 requires consistent estimation of the Q function. Assumption 6, Bellman completeness, is a standard structural restriction required for consistency.

Next we establish convergence rates of $\hat{\tau}^\pi$, depending on convergence rates of the nuisance functions. Broadly we follow the analysis of (Foster and Syrgkanis, 2019; Lewis and Syrgkanis, 2021) for orthogonal statistical learning. We show the loss function is Neyman-orthogonal in Appendix C.2. Given our high-level assumptions on product error rates, we state simplified results with "$\lesssim$" denoting $= O_p(\cdot)$ with high probability, omitting absolute multiplicative constants but for concentrability (assumption 3).

Theorem 4.1 (Policy evaluation, MSE rates)**.** *Suppose* $\left\{\sup_{s,t} \mathbb{E}[(A_t - \pi_t^b)(A_t - \pi_t^b) \mid S_t = s]\right\} \leq C$*, Assumptions 1 to 3, 5 and 6 and eqn.* (1)*, and product RMSE error rates for* π_b, Q *are* $o_p(n^{-\frac{1}{4}})$ *(Assumption 4). Fix the evaluation policy* π^e*. Then, for* $\sigma > 0$,

$$\|\hat{\tau}_t^{\pi^e} - \tau_t^{0,\pi^e}\|_2 \lesssim n^{-\frac{1}{2}}. \tag{6}$$

Our theorem states that estimation of π_b and Q^{π^e} needs to be only $n^{-\frac{1}{4}}$ convergent to guarantee that the product error rate of Assumption 4 holds with rate $n^{-\frac{1}{2}}$. Naive FQE would require $n^{-\frac{1}{2}}$ convergence of Q. Consistent estimation is still required for both Q and π^b functions, inheriting other conditions in the literature.

To compare to prior methods, in general, methods based on direct regression such as FQE will inherit sample complexity estimation *rates* from the empirical risk minimization subproblems. For example, (Duan et al., 2021) give a bound with Rademacher complexity on the Bellman error of $C \sum_{t=1}^T \mathcal{R}_n^{q_t}(\mathcal{F}_t) + O(n^{-1/2})$, where $\mathcal{R}_n^{q_t}(\mathcal{F}_t)$ is the Rademacher complexity of the function class $\mathcal{F}_t$, and $\mathcal{F}_t$ might be infinite or nonparametric. Nonparametric function classes are known to suffer curses of dimensionality. (Stone, 1980) proved that a p-smooth d-dimensional density function converges at a rate of at most $n^{-p/(2p+d)}$. For large d, nonparametric estimation is much slower, $p/(2p+d) \ll \frac{1}{2}$; hence complex nonparametric regression for Q, without sparsity or intrinsic dimension, may suffer rates slower than $n^{-1/2}$, which would slow down FQE rates. Where our results improve is that we can leverage estimation of π_b (which may be known in RL) in the orthogonal loss

function so that we only need $n^{-1/4}$-consistent estimation of both π^b, Q in order to achieve $n^{-1/2}$-consistent estimation of the difference-of-Q.

Policy optimization. Convergence of τ_t implies convergence in policy value. We quantify this with the *margin* assumption, a low-noise condition that quantifies the gap between regions of different optimal action (Tsybakov, 2004), often used to relate plug-in estimation error to decision risk.

Assumption 7 (Margin on observational distribution)**.** Let $Q_t^*(s, \pi^*)$ be the optimal Q function at the optimal action, and a' the second-best option, $a' \in \mathcal{A} \setminus \arg\max_a Q_t^*(s, a)$. Assume there exist constants α (the margin exponent), and $\delta_0 > 0$ such that

$$P(Q_t^*(S_t, \pi^*) - Q_t^*(S_t, a') \leq \epsilon) \leq (\varepsilon/\delta_0)^\alpha, \forall t \in 1, \ldots, T$$

The above probability is over the observational data distribution, $S_t \sim \mu_{\pi_t^b}$. Note that (Hu et al., 2024) establishes margin constants for linear and tabular MDPs; $\alpha = \infty$ for tabular MDPs, $\alpha = 1$ for linear MDPs where Q^* is linear, and for nonlinear Q^* under structural assumptions. A uniform gap between optimal and next-optimal action implies $\alpha = \infty$. The margin assumption allows us to translate advantage-function convergence rates to the policy value convergence rates, summarized in the following Lemma.

Lemma 1 (Advantage estimation error to policy value via margin.)**.** *Suppose Assumptions 1 to 3, 5 and 6 and eqn.* (1) *and Assumption 7 (margin assumption holds with α). Suppose that with high probability $\geq 1 - n^{-\kappa}$ for any finite $\kappa > 0$, $\|\hat{\tau}_t^n(s) - \tau_t^\circ(s)\|_2 \leq K\left(n^{-b_*}\right)$, for some rate $b_* > 0$. Then*

$$\left|\mathbb{E}[V_t^*(S_t) - V_t^{\hat{\pi}_{\hat{\tau}}}(S_t)]\right| \lesssim \frac{C_\infty}{1-\gamma} n^{-b_*\left(\frac{2+2\alpha}{2+\alpha}\right)}, \text{ and}$$
$$\|Q_t^*(S_t, \pi^*) - Q_t^*(S_t, \hat{\pi}_{\hat{\tau}})\|_2 \lesssim \frac{C_\infty}{1-\gamma} n^{-b_*\left(\frac{2+2\alpha}{2+\alpha}\right)}.$$

Next we study policy optimization. Lemma 1 relies on convergence of estimated difference-of-Q under the estimated-optimal policy to the true difference-of-Q under the optimal policy, while earlier guarantees in Theorem 4.1 just grant convergence for the estimated-optimal policy. This is the key technical difficulty that we overcome with induction and Assumption 7. We state a simplified version of the full theorem in the main text; the full statement is in Appendix B.2.

Theorem 4.2 (Policy optimization bound)**.** *Suppose Assumptions 1 to 6 and eqn.* (1) *and Assumption 7 (margin assumption holds with α). Suppose the product error rate conditions of Assumption 4 hold for each t for $\underline{\hat{\pi}}_{t+1}$, the data-optimal policies evaluated along the algorithm steps. Then for $\hat{\pi}_t$, Theorem 4.1 holds. And,*

$$\|\hat{\tau}_t^{\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ, \underline{\pi}^*_{t+1}}\|_2 \lesssim n^{-\frac{1}{2}},$$
$$\left|\mathbb{E}[V_1^{\pi^*}(S_1) - V_1^{\hat{\pi}_{\hat{\tau}}}(S_1)]\right| \lesssim C_\infty n^{-\{\frac{1}{2}\}\frac{2+2\alpha}{2+\alpha}}.$$

The main takeaway is that convergent policy evaluation admits convergent policy optimization, even though the nuisance functions evaluate an estimated-optimal, rather than population-optimal policy. Our method introduces auxiliary estimation at every timestep: although this *prevents faster* exponentiated convergence, it also *protects* the method from error accumulation from policy-dependent nuisance estimation.

Our Theorem 4.2 presents the main technical novelty in the analysis, whereas Theorem 4.1 follows existing approaches. The technical challenge arises from the policy-dependent nuisance, $Q_t^{\overline{\pi}_{t+1}}$, since it depends on the (optimized) policy at future timesteps. The margin condition shows that the error due to policy-dependent nuisances, under an estimated optimal policy, is higher-order relative to the time-t estimation of the advantage functions. The result follows by induction.

5 EXPERIMENTS

Adapting to structure in $\tau(s)$. Recent research highlights implications of blockwise conditional independence properties in RL, where some components are "exogenous" or irrelevant to rewards and actions (Wang et al., 2022; Wang, Xiao, Zhu, and Stone, Wang et al.; Dietterich et al., 2018). These methods may be designed for a particular graphical structure, and may be brittle under different substructures. Pretesting for the presence or absence of graphical restrictions incurs poor statistical properties. We advocate a different approach: by estimating the *difference-of-Q* functions, we can exploit statistical implications of underlying structure via sparse τ, without vulnerability to assumptions on the underlying data-generating process.

We investigate the benefits of estimating the difference-of-Qs in two different graphical substructures, Section 1, proposed in Zhou (2024b); Dietterich et al. (2018). Orthogonal causal contrast estimation is robust under noisy nuisance functions, illustrating our theory, and it can adapt to a variety of structures.

First we describe the modified Reward-Filtered DGP (left, Figure 2) of (Zhou, 2024b). We generate states of dimension $d = 100$, where only a small block (about 15 coordinates) is relevant for rewards and transitions, following the reward-filtered factorization of Zhou (2024b). Rewards depend on this sparse block plus additive noise, and transition matrices are block matrices that satisfy the graphical restrictions of Section 1. Thus

Table 2: Performance comparison across different sample sizes in nonlinear settings. Results show MSE ± SE.

Method	Number of Samples (n)				
	100	200	400	600	800
CartPole with Distractors					
FQE	2.367 ± 2.157	0.587 ± 0.772	1.157 ± 2.219	1.793 ± 1.618	4.123 ± 3.901
DiffQ	2.212 ± 2.376	0.415 ± 0.463	1.228 ± 1.831	1.929 ± 2.126	2.440 ± 1.912
DiffQ+MI	$\mathbf{2.104 \pm 2.392}$	$\mathbf{0.280 \pm 0.222}$	$\mathbf{1.179 \pm 1.840}$	$\mathbf{1.286 \pm 1.123}$	$\mathbf{2.342 \pm 1.812}$
Fourrooms Navigation ($\times 10^{-3}$)					
FQE	0.299 ± 0.520	0.194 ± 0.167	0.238 ± 0.230	0.175 ± 0.096	0.201 ± 0.151
DiffQ	0.281 ± 0.456	0.177 ± 0.142	0.224 ± 0.261	0.173 ± 0.055	0.187 ± 0.158
DiffQ+MI	$\mathbf{0.272 \pm 0.461}$	$\mathbf{0.173 \pm 0.171}$	$\mathbf{0.219 \pm 0.239}$	$\mathbf{0.170 \pm 0.094}$	$\mathbf{0.180 \pm 0.174}$

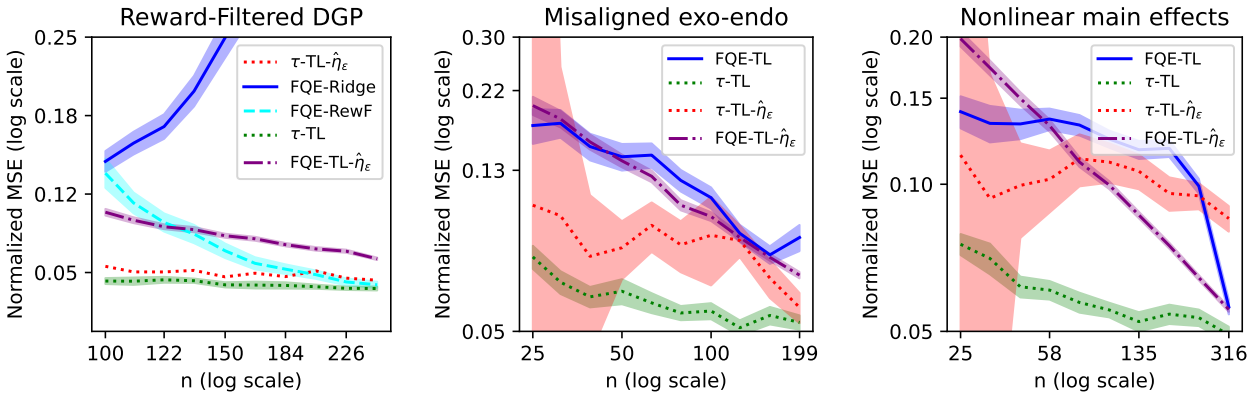


Figure 2: Adapting to structure. Interacted setting where $E[M_1 - M_0] = 0.1 \cdot I$.

the difference-of-Q function τ is sparse in the low-dimensional component, while naive Q estimation must contend with the full 100-dimensional state. Detailed generation (transition matrices, noise scales, policy specification) is deferred to Appendix E.

In Figure 2 we compare against baselines. In blue is FQE-Ridge, i.e. naive fitted-Q-evaluation with ridge regression. In dotted cyan is FQE-RF, the reward-filtered method of (Zhou, 2024b). Next we have two variants of our framework: in dotted green τ-TL which uses reward-based thresholding to estimate τ on the recovered support, and dotted-red τ-TL-$\hat{\eta}_\epsilon$, the same method with sample splitting with noisy nuisances. With τ-TL-$\hat{\eta}_\epsilon$, we investigate semi-synthetic settings with noisy nuisance functions by adding $N(0, n^{-1/4})$ noise to nuisance function predictions. For comparison to illustrate a setting with slow nuisance function convergence, we also include in dot-dashed purple FQE-TL-$\hat{\eta}_\epsilon$, which adds $n^{-1/4}$ noise to the oracle difference-of-Q function (estimated with LASSO). For our methods, we optimize with CVXPY.

We describe the results left to right. We display the mean over 100 replications (fixing the coefficient matrices and vectors, etc. with the same random seed); except for sample-splitting where we display the median. (With small n, sample splitting suffers finite-sample issues of small data splits, though this vanishes as n increases). The y-axis is the normalized MSE (divided by the squared range of true τ), and the x axis is the (log) number of episodes. First on the left, we consider the previously mentioned reward-filtered DGP. Zhou (2024a) is well-specified. For the reward-filtered DGP, we compare to FQE ridge regression, which we also use as a nuisance estimator for our approach. When compared to oracle-sparse difference-of-Q estimation, naive ridge FQE even diverges. However, our methods with thresholded LASSO do well, even if we plug-in the nuisance Q function estimated with Ridge regression. Orthogonal estimation is robust to nuisance function estimation error, as indicated by the red-dotted line where we plug-in quarter-root consistent estimates.

Next we slightly modify the graphical structure. Our methods adapt to the underlying sparsity in the difference-of-Q functions, *even if* the exact graphical independences differ. In all the experiments, naive

cross-validation does poorly. This is expected since cross-validation for predictive error doesn't ensure support recovery, unlike thresholded LASSO, and suffers extra challenges of hyperparameters in offline RL.

In "Misaligned endo-exo", we follow the same data-generating process as the "Reward-Filtered DGP" described earlier, but we change the blockwise conditional independences to follow the exogeneous-endogenous model of (Dietterich et al., 2018) (see Section 1). We additionally added dense rewards to the reward vector, adding $\beta_{dense}^\top \phi_t(s,a)$ where the entries of β_{dense} are 1 w.p. 0.9. Here, reward sparsity of $R(s,a), a \in \{0,1\}$ alone does not recover the sparse component. Reward-filtered thresholded LASSO is simply misspecified and does very poorly (off the graph limits). Likewise, in small samples, vanilla thresholded LASSO FQE (FQE-TL, dark-blue) includes too many extra dimensions. But for small-data regimes, imposing thresholded LASSO *on the difference of Q functions* remains stable.

The final DGP introduces "nonlinear main effects": again we generate a 50% dense vector β_{dense} and we add $s^\top \beta_{dense} + 3\sin(\pi s_{49} s_{48}) + 0.5(s_{49} - 0.5)^2 + 0.5(s_{48} - 0.5)^2$. (These nonlinear main effects are disjoint from the sparse difference-of-Q terms). For small n, FQE wrongly includes extraneous dimensions that destabilize estimation, and our methods estimating τ with reward-thresholded-LASSO outperform naive FQE with thresholded-LASSO for small data sizes.

Extending to nonlinear settings: mutual information regularization. Our experiments showcase that support recovery is necessary. To illustrate how the loss function approach permits more complex parametrizations, we now consider neural-nets and introduce a heuristic nonlinear sparsity-encouraging regularizer based on mutual information regularization. We seek a simpler representation that retains information related to the loss, while discarding irrelevant information unrelated to the proxy loss for the difference-of-Q functions. We use a mutual information regularizer (MIR), $\hat{\mathcal{L}}_{MI}(\phi,\theta)$, to encourage decomposing the state S into independent nonlinear representations X_c^ϕ, X_a^θ, parametrized respectively by ϕ, θ. Mutual information quantifies the dependency between two variables and it equals zero if and only if they are (marginally) independent. We parametrize the difference-of-Q function as $\tau(X_c^\phi)$, depending only on the *confounding* information X_c^ϕ which is relevant to the difference-of-Q loss function, while the *auxiliary* information X_a^θ is independent of the loss function. We also add a reconstruction loss function $\hat{\mathcal{L}}_{rec}(\phi,\theta)$ which ensures that these two representations jointly recover the state. These additional loss functions are weighted by hyperparameters λ_m, λ_r.

$$\hat{\mathcal{L}}^{nl}(\tau,\eta;\phi,\theta) = \hat{\mathcal{L}}(\tau_\phi,\eta) + \lambda_m \hat{\mathcal{L}}_{MI}(\phi,\theta) + \lambda_r \hat{\mathcal{L}}_{rec}(\phi,\theta),$$
$$\text{where } \hat{\mathcal{L}}_{MI} = |\widehat{I}(X_a^\phi; X_c^\theta)|, \hat{\mathcal{L}}_{rec} = \mathbb{E}[(X_a^\phi + X_c^\theta - S)^2]$$

Estimating mutual information is challenging. We use a recently developed mutual information neural-networks based estimator, abbreviated MINE (Belghazi et al., 2018). (See Appendix E for more details). MINE defines a neural information measure $I_\Lambda(X_a, X_c) = \sup_{\lambda\in\Lambda} {}_{X_a X_c}[\lambda] - \log({}_{X_a X_c}[e^\lambda])$. Usually MI requires access to probability densities, though only samples from the joint distribution in ML-based methods are available. MINE uses these samples.

Cartpole with Distractors results. We use a cartpole-with-distractors environment from Hao et al. (2024), which adds additional autoregressive noise to the state in CartPole[6] (Brockman et al., 2016). Hao et al. (2024) focuses on off-policy evaluation with abstractions, so the methods are not comparable. We illustrate how our method improves upon naive FQE (learned with neural nets) learned on the full state space. The results are in Table 2. We evaluate the MSE relative to an oracle difference-of-Q function obtained by differencing Q estimates from FQE from a large dataset, $n = 2000$, trained only on the original 4-d state space. We compare to our DiffQ estimation with neural nets, and a regularized version.

Four Rooms Environment Results. To further validate our approach in complex discrete environments, we evaluated our method on the D4RL/minigrid fourrooms-random-v0 environment, which features a discrete action space with 7 actions and high-dimensional state representations (148 dimensions from visual input). This environment requires complex spatial reasoning and navigation. The Four Rooms results in Table 2 confirm that our method successfully *scales* to environments with multiple discrete actions and high-dimensional state representations. Further including irrelevant distractors could yield further improvements.

Conclusions and limitations. We bridged offline RL and causal ML by developing orthogonal methodology for estimating the difference-of-Q functions. Improving policy optimization with standard tools like pessimism, and weakening the concentrability assumption, are important for future work; we expect standard techniques in the literature can help here. Though our policy learning relies on a margin assumption, similar assumptions are required for sparse learning methods. Therefore, our approach would not work well in settings with very many small nonzero coefficients.

[6] Given this environment, a long finite-horizon environment with time-homogenous transitions, we learn this as a $\gamma = 0.99$ discounted infinite-horizon problem and pool the data. In the appendix we discuss the extension.

References

Belghazi, M. I., A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm (2018). Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR.

Brockman, G., V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

Chakraborty, B. and S. A. Murphy (2014). Dynamic treatment regimes. *Annual review of statistics and its application 1*, 447–464.

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

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

Chernozhukov, V., C. Hansen, N. Kallus, M. Spindler, and V. Syrgkanis (2024). Applied causal inference powered by ml and ai. *rem 12*(1), 338.

Dietterich, T., G. Trimponias, and Z. Chen (2018). Discovering and removing exogenous state variables and rewards for reinforcement learning. In *International Conference on Machine Learning*, pp. 1262–1270. PMLR.

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

Ernst, D., G.-B. Stan, J. Goncalves, and L. Wehenkel (2006). Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 667–672. IEEE.

Farias, V., A. Li, T. Peng, and A. Zheng (2022). Markovian interference in experiments. *Advances in Neural Information Processing Systems 35*, 535–549.

Foster, D. J., A. Krishnamurthy, D. Simchi-Levi, and Y. Xu (2021). Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*.

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

Gupta, G., T. G. J. Rudner, R. T. McAllister, A. Gaidon, and Y. Gal (2023, 11–14 Apr). Can active sampling reduce causal confusion in offline reinforcement learning? In M. van der Schaar, C. Zhang, and D. Janzing (Eds.), *Proceedings of the Second Conference on Causal Learning and Reasoning*, Volume 213 of *Proceedings of Machine Learning Research*, pp. 386–407. PMLR.

Hao, M., P. Su, L. Hu, Z. Szabo, Q. Zhao, and C. Shi (2024). Forward and backward state abstractions for off-policy evaluation. *arXiv preprint arXiv:2406.19531*.

Hu, Y., N. Kallus, and M. Uehara (2024). Fast rates for the regret of offline reinforcement learning. *Mathematics of Operations Research*.

Jayaraman, P., J. Desman, M. Sabounchi, G. N. Nadkarni, and A. Sakhuja (2024). A primer on reinforcement learning in medicine for clinicians. *NPJ Digital Medicine 7*(1), 337.

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

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

Kallus, N. and M. Uehara (2019a). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *arXiv preprint arXiv:1908.08526*.

Kallus, N. and M. Uehara (2019b). Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*.

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

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

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences 116*(10), 4156–4165.

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

Levine, S., A. Kumar, G. Tucker, and J. Fu (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

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

Lewis, G. and V. Syrgkanis (2021). Double/debiased machine learning for dynamic treatment effects. *Advances in Neural Information Processing Systems 34*.

Li, F., L. E. Thomas, and F. Li (2019). Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology 188*(1), 250–257.

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

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. (2015). Human-level control through deep reinforcement learning. *nature 518*(7540), 529–533.

Morzywolek, P., J. Decruyenaere, and S. Vansteelandt (2023). On a general class of orthogonal learners for the estimation of heterogeneous treatment effects. *arXiv preprint arXiv:2303.12687*.

Munos, R. and C. Szepesvári (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research 9*(5).

Murphy, S. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology 65*(2), 331–355.

Neumann, G. and J. Peters (2008). Fitted q-iteration by advantage weighted regression. *Advances in neural information processing systems 21*.

Nie, X., E. Brunskill, and S. Wager (2021). Learning when-to-treat policies. *Journal of the American Statistical Association 116*(533), 392–409.

Nie, X. and S. Wager (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika 108*(2), 299–319.

Pan, H.-R. and B. Schölkopf (2023). Learning endogenous representation in reinforcement learning via advantage estimation. In *Causal Representation Learning Workshop at NeurIPS 2023*.

Pan, H.-R. and B. Schölkopf (2024). Skill or luck? return decomposition via advantage functions. *arXiv preprint arXiv:2402.12874*.

Rotnitzky, A., E. Smucler, and J. M. Robins (2021). Characterization of parameters with a mixed bias property. *Biometrika 108*(1), 231–238.

Schulte, P. J., A. A. Tsiatis, E. B. Laber, and M. Davidian (2014). Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the Institute of Mathematical Statistics 29*(4), 640.

Shi, C., S. Luo, H. Zhu, and R. Song (2022). Statistically efficient advantage learning for offline reinforcement learning in infinite horizons. *arXiv preprint arXiv:2202.13163*.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, 1348–1360.

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

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics 32*(1), 135–166.

Voloshin, C., H. M. Le, N. Jiang, and Y. Yue (2019). Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

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

Wang, T., S. S. Du, A. Torralba, P. Isola, A. Zhang, and Y. Tian (2022). Denoised mdps: Learning world models better than the world itself. *arXiv preprint arXiv:2206.15477*.

Wang, Z., X. Xiao, Y. Zhu, and P. Stone. Task-independent causal state abstraction.

Xie, C., W. Yang, and Z. Zhang (2023). Semiparametrically efficient off-policy evaluation in linear markov decision processes. In *International Conference on Machine Learning*, pp. 38227–38257. PMLR.

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

Xu, T., Z. Yang, Z. Wang, and Y. Liang (2021). Doubly robust off-policy actor-critic: Convergence and optimality. In *International Conference on Machine Learning*, pp. 11581–11591. PMLR.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika 100*(3), 681–694.

Zhou, A. (2024a). Optimal and fair encouragement policy evaluation and learning. *Advances in Neural Information Processing Systems 36*.

Zhou, A. (2024b). Reward-relevance-filtered linear offline reinforcement learning.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, in Section 3
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, in Section 4, primarily on statistical complexity.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes, in Section 4 in the main text we include simplified statements of the theorems, and full statements of the required assumptions. Full theorem statements are in Appendix B.2.
   (b) Complete proofs of all theoretical results. Yes, but all proofs are in the appendix, Appendix C.
   (c) Clear explanations of any assumptions. Yes, in the main text, right after the assumption statement in Section 4.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, in appendix.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, in Appendix E.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Additional discussion

Additional discussion on related work. Pan and Schölkopf (2023) note the analogy of the advantage function with causal contrast estimation and derive a Q-function independent estimator, but in the online setting. After preparing an initial version of this paper, we became aware of the recent work of Pan and Schölkopf (2024) in the offline setting. While their method elegantly avoids estimating future Q functions, it requires a nonconvex constraint on the action-average of advantages, which is computationally and statistically difficult. The motivations are different; the methods are complementary. (See appendix for more discussion.) Our identification is different and we focus on improved statistical guarantees.

Extension to stationary discounted infinite-horizon setting. The identification argument extends to the stationary discounted infinite-horizon setting. For policy optimization, we make a small modification: instead of iterative optimization and estimation, we first conduct offline policy optimization to estimate the optimal policy and its Q function, $\hat{\pi}'$ and $\hat{Q}^{\hat{\pi}'}$. We describe the algorithm in Algorithm 2. This can be done with a variety of methods that are common and popular in practice, such as DQN (Mnih et al., 2015), fitted-Q-iteration (Chen and Jiang, 2019), or other algorithms.

Algorithm 2 Stationary Infinite-Horizon Dynamic Residualized Difference-of-Q Optimization

1: Given: Partition of $\mathcal{D}$ into 3 folds, $\{\mathcal{D}_k\}_{k=1}^3$.
2: Estimate $\hat{\pi}_t^b$ on $\mathcal{D}_1$.
3: Estimate $\hat{Q}^{\hat{\pi}'}$ on $\mathcal{D}_1$ with offline policy optimization. Evaluate $m_t^{\hat{\pi}'}$.
4: Estimate $\hat{\tau}_t^{\hat{\pi}'}$ on $\mathcal{D}_{k(t)}$ by minimizing the empirical loss:

$$\hat{\tau}_t(\cdot) \in \operatorname{argmin}_\tau \sum_{i \in \mathcal{D}_{k(t)}} \left(R_t^i + \gamma \sum_{a'} \hat{\pi}_{t+1}\left(a' \mid S_{t+1}^i\right) \hat{Q}^{\hat{\pi},(1)}\left(S_{t+1}^i, a'\right) - \hat{m}_t^{\hat{\pi}'}(S_t^i) - \{A_t^i - \hat{\pi}_t^{b,(1)}(S_t^i)\}\tau_t(S_t^i) \right)^2$$

5: Policy optimization: For two actions, $\hat{\pi}_t(s) = \mathbb{I}[\hat{\tau}_t^{\hat{\pi}'}(s) > 0]$. Else for multiple actions, $\hat{\pi}_t(s) \in \arg\max_{a' \in \mathcal{A} \setminus a_0} \hat{\tau}_t^{\hat{\pi}'}(s, a')$ if $\max_{a' \in \mathcal{A} \setminus a_0} \hat{\tau}_t^{\hat{\pi}'}(s, a') > 0$, else a_0.

Algorithm 3 Dynamic Residualized Difference-of-Q-Evaluation

Given: π^e, evaluation policy; and for sample splitting, partition of $\mathcal{D}$ into K folds, $\{\mathcal{D}_k\}_{k=1}^K$.
On $\mathcal{D}_k$, estimate $\hat{Q}^{\pi^e,k}$, behavior policy $\hat{\pi}_t^{b,k}$, therefore $\hat{m}^{\pi^e,k}$.
for timestep $t = T, \ldots, 1$ **do**
$\hat{\tau}_t \in \text{argmin}_\tau \{ \sum_{k=1}^K \sum_{i \in \mathcal{D}_k} \left(R_t^i + \gamma \hat{Q}_{t+1}^{\hat{\pi},-k}(S_{t+1}^i, A_{t+1}^i) - \hat{m}_t^{\hat{\pi},-k}(S_t^i) - \{A_t^i - \hat{\pi}_t^{b,-k}(S_t^i)\}\tau_t(S_t^i) \right)^2 \}$
end for

B Deferred full statements

B.1 Full statements of algorithms

B.2 Full statements of theorems

Remark 1 (Simplified theorems in main instantiate these general theorem statements). In the main text, we present simplified theorem statements for readability and to convey the essential results. Here we include the more general statements for completeness. Theorems 4.1 and 4.2 can be obtained via Assumption 5 (well-specification and exact solution so that $\epsilon(\tau_t^n, \hat{\eta}) = 0$, and $\tau_t^{\pi^e,n} = \tau_t^{\pi^e,\circ}$ below. In the main text we assume that π_b, Q^π are $o_p(n^{-\frac{1}{4}})$ RMSE-consistent, whereas these are quantified via the critical radius $\delta_{n/2}^2$ term below.

The analysis considers some generic candidate $\hat{\tau}$ with small excess risk relative to the projection onto the function class, i.e. as might arise from an optimization algorithm with some approximation error. For a fixed evaluation policy π^e, define the projection of the true advantage function onto Ψ^n, $\tau_t^{\pi^e,n} = \arg\inf_{\tau_t \in \Psi_t^n} \|\tau_t - \tau_t^{\circ,\pi^e}\|_2$, and the error $\nu_t^{\pi^e} = \hat{\tau}_t^{\pi^e} - \tau_t^{n,\pi^e}$ of some estimate $\hat{\tau}_t^{\pi^e}$ to projection onto the function class:

Theorem B.1 (Policy evaluation). *Suppose* $\left\{\sup_{s,t} \mathbb{E}[(A_t - \pi_t^b)(A_t - \pi_t^b) \mid s]\right\} \leq C$ *and Assumptions 1 and 2 and eqn.* (1). *Consider a fixed evaluation policy* π^e. *Consider any estimation algorithm that produces an estimate* $\hat{\tau}^{\pi^e} = \left(\tau_1^{\pi^e}, \ldots, \tau_T^{\pi^e}\right)$, *with small plug-in excess risk at every* t, *with respect to any generic candidate* $\tilde{\tau}^{\pi^e}$, *at some nuisance estimate* $\hat{\eta}$, *i.e.,*

$$\mathcal{L}_{D,t}(\hat{\tau}_t^{\pi^e}; \hat{\eta}) - \mathcal{L}_{D,t}(\tilde{\tau}_t^{\pi^e}; \hat{\eta}) \leq \epsilon(\tau_t^n, \hat{\eta}).$$

Let ρ_t *denote product error terms:*

$$\begin{aligned}\rho_t^{\pi^e}(\hat{\eta}) = {B_\tau}^2 \|(\hat{\pi}_t^b - \pi_t^{b,\circ})^2\|_u + B_\tau \|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{m}_t^{\pi^e} - m_t^{\pi^e,\circ})\|_u \\ + \gamma(B_\tau \|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{Q}_{t+1}^{\pi^e} - Q_{t+1}^{\pi^e,\circ})\|_u + \|(\hat{m}_t^{\pi^e} - m_t^{\pi^e,\circ})(\hat{Q}_{t+1}^{\pi^e} - Q_{t+1}^{\pi^e,\circ})\|_u).\end{aligned} \tag{7}$$

Then, for $\sigma > 0$, *and* $u^{-1} + \overline{u}^{-1} = 1$,

$$\frac{\lambda}{2}\|\nu_t^{\pi^e}\|_2^2 - \frac{\sigma}{4}\|\nu_t^{\pi^e}\|_{\overline{u}}^2 \leq \epsilon(\hat{\tau}_t^{\pi^e}, \hat{\eta}) + \frac{2}{\sigma}\left(\|(\tau^{\pi^e,\circ} - \tau_t^{\pi^e,n})\|_u^2 + \rho_t^{\pi^e}(\hat{\eta})^2\right).$$

In the above theorem, $\epsilon(\hat{\tau}_t^{\pi^e}, \hat{\eta})$ is the excess risk of the empirically optimal solution. Note that in our setting, this excess risk will be an approximation error incurred from the proxy loss issue described in Lemma 3.

The bias term is $\|(\tau^{\pi^e,\circ} - \tau_t^{\pi^e,n})\|_u^2$, which describes the model misspecification bias of the function class parametrizing $Q-$function contrasts, Ψ.

The product error terms $\rho_t^{\pi^e}(\hat{\eta})$ highlight the reduced dependence on individual nuisance error rates. We will instantiate the previous generic theorem for the projection onto Ψ^n, $\tau_t^{\pi^e,n}$, also accounting for the sample splitting. We will state the results with *local Rademacher complexity*, which we now introduce. For generic 1-bounded functions f in a function space $f \in \mathcal{F}, f \in [-1, 1]$, the local Rademacher complexity is defined as follows:

$$\mathscr{R}_n(\mathcal{F}; \delta) = \mathbb{E}_{\epsilon_{1:n}, X_{1:n}}\left[\sup_{f \in \mathcal{F}: \|f\|_2 \leq \delta} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right]$$

The critical radius δ^2 more tightly quantifies the statistical complexity of a function class, and is any solution to the so-called *basic inequality*, $\mathscr{R}_n(\mathcal{F}; \delta) \leq \delta^2$. The star hull of a generic function class $\mathcal{F}$ is defined as

star$(\mathcal{F}) = \{cf : f \in \mathcal{F}, c \in [0,1]\}$. Bounds on the critical radius of common function classes like linear and polynomial models, deep neural networks, etc. can be found in standard references on statistical learning theory, e.g. (Wainwright, 2019). We can obtain mean-squared error rates for policy evaluation via specializing Theorem 4.1 to the 2-norm and leveraging results from (Foster and Syrgkanis, 2019).

Assumption 8 (Product error rates on nuisance function evaluation)**.** Fix an evaluation policy π^e. Suppose each of $\mathbb{E}[\|(\hat{\pi}_t^b - \pi_t^{b,\circ})\|_2^2]$, $\mathbb{E}[\|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{m}_t^{\pi^e} - m_t^{\pi^e,\circ})\|_2^2]$, $\mathbb{E}[\|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{Q}_{t+1}^{\pi^e} - Q_{t+1}^{\pi^e,\circ})\|_2^2]$, and $\mathbb{E}[\|(\hat{m}_t^{\pi^e} - m_t^{\pi^e,\circ})(\hat{Q}_{t+1}^{\pi^e} - Q_{t+1}^{\pi^e,\circ})\|_2^2]$ are of order $O(\delta_{n/2}^2 + \|\tau_t^{\pi^e,\circ} - \tau_t^{\pi^e,n}\|_2^2)$.

Denote the leading order of the product-error rate as ρ, i.e. $O(\delta_{n/2}^2 + \|\tau_t^{\pi^e,\circ} - \tau_t^{\pi^e,n}\|_2^2) = O(n^{-\rho})$.

Assumption 4 summarizes both the product-error estimation rates and the misspecification error for τ, $\|\tau_t^{\pi^e,\circ} - \tau_t^{\pi^e,n}\|_2^2$, in the rate term ρ. Meanwhile, Assumption 4 requires consistent estimation of the Q function. Therefore we inherit potentially stringent structural restrictions for Q-function estimation such as Bellman completeness or linear Bellman completeness (Foster et al., 2021). Our orthogonal estimation enjoys the so-called "rate-double robustness" property, i.e. requiring only product-error $n^{-\frac{1}{2}}$ convergence of Q, π_b, but not the "mixed-bias" property of double-robustness wherein only one of Q or π_b need to be well-specified for unbiased estimation (Rotnitzky et al., 2021).

Theorem B.2 (MSE rates for policy evaluation)**.** *Suppose* $\left\{\sup_{s,t} \mathbb{E}[(A_t - \pi_t^b)(A_t - \pi_t^b) \mid s]\right\} \leq C$ *and Assumptions 1, 2 and 9 and eqn.* (1)*. Suppose Assumption 4 holds with rate* ρ*, i.e. Consider a fixed policy* π^e*. Then*

$$\mathbb{E}[\|\hat{\tau}_t^{\pi^e} - \tau_t^{\pi^e,\circ}\|_2^2] = O\left(\delta_{n/2}^2 + \|\tau_t^{\pi^e,\circ} - \tau_t^{\pi^e,n}\|_2^2\right)$$

Under stronger assumptions on the MDP, we can provide stronger sup-norm convergence guarantees on the difference-of-Q. In the main text, we make a slightly weaker concentrability (Assumption 3) for weaker convergence results in integrated risk.

Assumption 9 (Bounded transition density)**.** Transitions have bounded density: $P(s' \mid s, a) \leq c$. Let $d_\pi(s)$ denote the marginal state distribution under policy π. Assume that $d_{\pi_t^b}(s) < c$, for $t = 1, \ldots, T$.

Lemma 2 (Advantage estimation error to policy value via margin.)**.** *Suppose eqn.* (1) *and assumption 7 (margin assumption holds with* α*).*

Suppose Assumption 9. Suppose that with high probability $\geq 1 - n^{-\kappa}$ *for any finite* $\kappa > 0$*, the following sup-norm convergence holds with some rate* $b_* > 0$*,*

$\sup_{s \in \mathcal{S}, a \in \mathcal{A}} |\hat{\tau}_t^{\hat{\underline{\pi}}_{t+1}}(s) - \tau_t^{\underline{\pi}_{t+1}^*,\circ}(s)| \leq K n^{-b*}$,

then $\left|\mathbb{E}[V_t^*(S_t) - V_t^{\hat{\pi}_{\hat{\tau}}}(S_t)]\right| \leq \frac{(1-\gamma^{T-t})}{1-\gamma} c K^2 n^{-b*(1+\alpha)} + O(n^{-\kappa})$,

and $\|Q_t^*(S_t, \pi^*) - Q_t^*(S_t, \hat{\pi}_{\hat{\tau}})\|_2 \leq \frac{(1-\gamma^{T-t})}{1-\gamma} c K^2 n^{-b*(1+\alpha)} + O(n^{-\kappa})$.

Else, assume sup-norm concentrability (Assumption 3) and that $\|\hat{\tau}_t^n(s) - \tau_t^\circ(s)\|_2 \leq K\left(n^{-b_*}\right)$*, for some rate* $b_* > 0$*. Then*

$$\left|\mathbb{E}[V_t^*(S_t) - V_t^{\hat{\pi}_{\hat{\tau}}}(S_t)]\right| \lesssim \frac{(1-\gamma^{T-t})}{1-\gamma} C_\infty n^{-b_*\left(\frac{2+2\alpha}{2+\alpha}\right)}, \text{ and } \|Q_t^*(S_t, \pi^*) - Q_t^*(S_t, \hat{\pi}_{\hat{\tau}})\|_2 \lesssim \frac{(1-\gamma^{T-t})}{1-\gamma} C_\infty n^{-b_*\left(\frac{2+2\alpha}{2+\alpha}\right)}.$$

Theorem B.3 (Policy optimization bound)**.** *Suppose Assumptions 1, 2 and 9 and eqn.* (1)*. Further, suppose that* Q° *satisfies Assumption 7 (margin) with* $\alpha > 0$*. Suppose the product error rate conditions of Assumption 4 hold for each* t *for for* $\hat{\underline{\pi}}_{t+1}$*, the data-optimal policies evaluated along the algorithm steps, with rates* ρ_t *for each timestep. Suppose that then for* $\hat{\pi}_t$*, Theorem B.2 holds. Denote the slowest such product-error-rate over timesteps as* $\overline{\rho}_{1:T} = \min_t\{\rho_t\}$*. Then,*

$$\|\hat{\tau}_t^{\hat{\underline{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\pi}_{t+1}^*}\|_2 \leq O(n^{-\overline{\rho}_{1:T}}), \text{ and } \left|\mathbb{E}[V_1^{\pi^*}(S_1) - V_1^{\hat{\pi}_{\hat{\tau}}}(S_1)]\right| = O(n^{-\{\overline{\rho}_{1:T}\}\frac{2+2\alpha}{2+\alpha}}). \tag{8}$$

C Proofs

C.1 Preliminaries

Lemma 3 (Excess Variance).

$$\mathbb{E}[\hat{\mathcal{L}}_t(\tau,\eta)] - \mathcal{L}_t(\tau,\eta) = \mathrm{Var}[\max_{a'} Q\left(S_{t+1}, a'\right) \mid \pi_t^b]$$

Proof.

$$\begin{aligned}
&\mathbb{E}[\hat{\mathcal{L}}_t(\tau,\eta)] \\
&= \mathbb{E}[\Big(\{R_t + \gamma Q_{t+1}^{\pi^e}(S_{t+1}, A_{t+1}) - m_t^{\pi}(S_t)\} \pm \mathbb{E}[\gamma Q_{t+1}^{\pi^e} \mid S_t, \pi_t^b] - \{A - \pi_t^b(1 \mid S_t)\} \cdot \tau(S_t)\Big)^2] \\
&= \mathbb{E}[\Big(\{R_t + \gamma \mathbb{E}[Q_{t+1}^{\pi^e} \mid S_t, \pi_t^b] - m_t^{\pi}(S_t)\} - \{A - \pi_t^b(1 \mid S_t)\} \cdot \tau(S_t) + \gamma(Q_{t+1}^{\pi^e}(S_{t+1}, A_{t+1}) - \mathbb{E}[Q_{t+1}^{\pi^e} \mid S_t, \pi_t^b])\Big)^2] \\
&= \mathbb{E}[\Big(\{R_t + \gamma \mathcal{T} Q_{t+1}^{\pi^e} - m_t^{\pi}\} - \{A - \pi_t^b(1 \mid S_t)\} \cdot \tau(S_t)\Big)^2] && \text{(squared loss of identifying moment)} \\
&\quad + \mathbb{E}[\gamma(Q_{t+1}^{\pi^e}(S_{t+1}, A_{t+1}) - \mathbb{E}[Q_{t+1}^{\pi^e} \mid S_t, \pi_t^b])^2] && \text{(residual variance of } Q_t(s,a) - R_t(s,a)) \\
&\quad + \mathbb{E}[\Big\{R_t + \gamma \mathbb{E}[Q_{t+1}^{\pi^e} \mid S_t, \pi_t^b] - m_t^{\pi}(S_t) - \{A - \pi_t^b(1 \mid S_t)\} \cdot \tau(S_t)\Big\} \cdot \gamma(Q_{t+1}^{\pi^e}(S_{t+1}, A_{t+1}) - \mathbb{E}[Q_{t+1}^{\pi^e} \mid S_t, \pi_t^b])]
\end{aligned}$$

Note the last term $= 0$ by iterated expectations and the pull-out property of conditional expectation. □

C.2 Orthogonality

Below we will omit the π superscript; the analysis below holds for any valid π. Define $\nu_t = \hat{\tau}_t - \tau_t^n, \nu_t^\circ = \hat{\tau}_t - \tau_t^\circ$. We define for any functional $L(f)$ the Frechet derivative as:

$$D_f L(f)[\nu] = \frac{\partial}{\partial t} L(f + t\nu)\Big|_{t=0}$$

Higher order derivatives are denoted as $D_{g,f} L(f,g)[\mu,\nu]$.

Lemma 4 (Universal Orthogonality).

$$D_{\eta,\tau_t} \mathcal{L}_t(\tau_t^n; \tau_{t+1}^n, \eta^*)[\eta - \eta^*, \nu_t] = 0$$

Proof of Lemma 4. For brevity, for a generic f, let $\{f\}_\epsilon$ denote $f + \epsilon(f - f^\circ)$. Then the first Frechet derivatives are:

$$\frac{d}{d_{\epsilon_\tau}} \mathcal{L}_t(\tilde{\tau}, \eta^\circ)[\tau - \tilde{\tau}, \eta - \eta^\circ] = \mathbb{E}\left[\Big\{R_t + \gamma\{Q_{t+1}^{\pi^e,\circ}\}_\epsilon - \{m_t^{\pi^e,\circ}\}_\epsilon - (A_t - \{\pi_t^{b\circ}\}_\epsilon)\tau\Big\}(A_t - \{\pi_t^{b,\circ}\}_\epsilon)(\tilde{\tau} - \tau)\right]$$

$$\begin{aligned}
&\frac{d}{d\epsilon_e}\frac{d}{d\epsilon_\tau} \mathcal{L}_t\left(\tilde{\tau}, \eta^\circ\right)[\eta - \eta^\circ, \tau - \tau]\Big|_{\epsilon=0} \\
&\quad = \mathbb{E}\left[\left(\pi_t^b - \pi_t^{b,\circ}\right)\tau(\tau - \tilde{\tau})(A_t - e_t)\right] + \mathbb{E}\left[\Big\{R + \gamma Q_{t+1}^{\pi^e} - m_t^{\pi^e,\circ} - (A_t - e_t)\Big\}(\tau - \tilde{\tau}) \cdot -(e_t - e_t^\circ)\right] \\
&\quad = 0 \\
&\frac{d}{d\epsilon_{Q_{t+1}}}\frac{d}{d\epsilon_\tau} \mathcal{L}_t\left(\tilde{\tau}, \eta^\circ\right)[\eta - \eta^\circ, \tau - \tilde{\tau}]\Big|_{\epsilon=0} \\
&\quad = \mathbb{E}[\gamma(Q_{t+1}^{\pi^e} - Q_{t+1}^{\pi^e,\circ})(A_t - \pi_t^{b,\circ})(\tau_t - \tilde{\tau}_t)]] \\
&\quad = 0 \\
&\frac{d}{d\epsilon_{m_t}}\frac{d}{d\epsilon_\tau} \mathcal{L}_t\left(\tilde{\tau}, \eta^\circ\right)[\eta - \eta^\circ, \tau - \tilde{\tau}]\Big|_{\epsilon=0} \\
&\quad = \mathbb{E}[-(m_t^{\pi^e} - m_t^{\pi^e,\circ})(A_t - \pi_t^{b,\circ})(\tau_t - \tilde{\tau}_t)] \\
&\quad = 0
\end{aligned}$$

□

Lemma 5 (Second order derivatives)**.** *For Q_{t+1}, Q°_{t+1} evaluated at some fixed policy π^e:*

$$\begin{aligned}
&D_{\eta_t,\eta_t}\mathcal{L}_t[\hat{\eta}_t - \eta^\circ_t, \hat{\eta}_t - \eta^\circ_t] \\
&\quad = \mathbb{E}\left[\tau_t^2\left(\hat{\pi}^b_t - \pi^{b,\circ}_t\right)^2\right] + \mathbb{E}\left[(\hat{\pi}^b_t - \pi^{b,\circ}_t)\tau_t(\hat{m}_t - m^\circ_t)\right] + \mathbb{E}\left[(\hat{\pi}^b_t - \pi^{b,\circ}_t)\tau_t\gamma(\hat{Q}_{t+1} - Q^\circ_{t+1})\right] \\
&\quad - \mathbb{E}\left[(\hat{m}_t - m^\circ_t)\,\gamma\left(\hat{Q}_{t+1} - Q^\circ_{t+1}\right)\right]
\end{aligned}$$

Proof of Lemma 5. Below, the evaluation policy π^e is fixed and omitted for brevity. Note that

$$\begin{aligned}
D_e\mathcal{L}_D[\hat{e} - e^\circ] &= \mathbb{E}[(R_t + \gamma Q_{t+1} - \pi^{b\top}Q_t + (A - \pi^b_t)\tau_t)(-\tau_t)(\hat{e} - e^\circ)] \\
D_{m_t}\mathcal{L}_D[\hat{m}_t - m^\circ_t] &= \mathbb{E}[(R_t + \gamma Q_{t+1} - \pi^{b\top}Q_t + (A - \pi^b_t)\tau_t)(-1) * (m_t - m^\circ)]
\end{aligned}$$

By inspection, note that the nonzero terms of the second-order derivatives are as follows:

$$\begin{aligned}
D_{\pi^b_t,\pi^b_t}\mathcal{L}_t[\hat{\pi}^b_t - \pi^{b,\circ}_t, \hat{\pi}^b_t - \pi^{b,\circ}_t] &= \mathbb{E}\left[\tau_t^2\left(\hat{\pi}^b_t - \pi^{b,\circ}_t\right)^2\right] \\
D_{m_t,Q_{t+1}}\mathcal{L}_t[\hat{Q}_{t+1} - Q^\circ_{t+1}, \hat{m}_t - m^\circ_t] &= \mathbb{E}\left[-(\hat{m}_t - m^\circ_t)\,\gamma\left(\hat{Q}_{t+1} - Q^\circ_{t+1}\right)\right] \\
D_{m_t,\pi^b_t}\mathcal{L}_t[\hat{\pi}^b_t - \pi^{b,\circ}_t, \hat{m}_t - m^\circ_t] &= \mathbb{E}\left[(\hat{\pi}^b_t - \pi^{b,\circ}_t)\tau_t(\hat{m}_t - m^\circ_t)\right] \\
D_{Q_{t+1},\pi^b_t}\mathcal{L}_t[\hat{\pi}^b_t - \pi^{b,\circ}_t, \hat{Q}_{t+1} - Q^\circ_{t+1}] &= \mathbb{E}\left[(\hat{\pi}^b_t - \pi^{b,\circ}_t)\tau_t\gamma(\hat{Q}_{t+1} - Q^\circ_{t+1})\right]
\end{aligned}$$

By the chain rule for Frechet differentiation, we have that

$$\begin{aligned}
&D_{\eta_t,\eta_t}\mathcal{L}_t[\hat{\eta}_t - \eta^\circ_t, \hat{\eta}_t - \eta^\circ_t] = D_{\pi^b_t,\pi^b_t}\mathcal{L}_t[\hat{\pi}^b_t - \pi^{b,\circ}_t, \hat{\pi}^b_t - \pi^{b,\circ}_t] \\
&\quad + D_{m_t,\pi^b_t}\mathcal{L}_t[\hat{\pi}^b_t - \pi^{b,\circ}_t, \hat{m}_t - m^\circ_t] + D_{Q_{t+1},\pi^b_t}\mathcal{L}_t[\hat{\pi}^b_t - \pi^{b,\circ}_t, \hat{Q}_{t+1} - Q^\circ_{t+1}] + D_{m_t,Q_{t+1}}\mathcal{L}_t[\hat{Q}_{t+1} - Q^\circ_{t+1}, \hat{m}_t - m^\circ_t]
\end{aligned}$$

□

C.3 Proof of sample complexity bounds

Proof of Lemma 1. We begin with assuming the stronger assumption of Assumption 9, before discussing the weaker assumption of Assumption 3 and corresponding integrated risk bounds.

First, we use the following decomposition regardless of which concentrability-type assumption we use (Assumption 9 or Assumption 3).

$$\begin{aligned}
V^*_t(s) - V^{\pi_{\hat{\tau}}}_t(s) &= V^*_t(s) - V^{\pi_{\hat{\tau}}}_t(s) \pm Q^{\pi^*}(s, \pi_{\hat{\tau}}) \\
&= Q^*_t(s, \pi^*(s)) - Q^*_t(s, \hat{\pi}_{\hat{\tau}}) + Q^*_t(s, \hat{\pi}_{\hat{\tau}}) - V^{\hat{\pi}_{\hat{\tau}}}_t(s) \\
&\le \gamma\mathbb{E}_{\hat{\pi}_t}\left[V^{\pi^*}_{t+1} - V^{\hat{\pi}_{\hat{\tau}}}_{t+1} \mid s\right] + Q^*_t(s, \pi^*(s)) - Q^*_t(s, \hat{\pi}_{\hat{\tau}})
\end{aligned}$$

Therefore for any t and Markovian policy π inducing a marginal state distribution:

$$\mathbb{E}[V^*_t(s)] - \mathbb{E}[V^{\pi_{\hat{\tau}}}_t(s)] \le \gamma\mathbb{E}\left[\mathbb{E}_{\hat{\pi}_t}[V^{\pi^*}_{t+1} - V^{\hat{\pi}_{\hat{\tau}}}_{t+1} \mid s]\right] + \mathbb{E}[Q^*_t(s, \pi^*) - Q^*_t(s, \hat{\pi}_{\hat{\tau}})] \tag{9}$$

Assuming bounded rewards and Assumption 9 implies that $P(s_{t+1} \mid s, a) \le c$, which remains true under the state-action distribution induced by any Markovian policy $\pi(s, a)$, including the optimal policy. Therefore the second term of the above satisfies:

$$\mathbb{E}_\pi[Q^*_t(s_t, \pi^*) - Q^*_t(s_t, \hat{\pi}_{\hat{\tau}})] \le c\int \{Q^*_t(s, \pi^*) - Q^*_t(s, \hat{\pi}_{\hat{\tau}}\}\, ds, \tag{10}$$

and fixing $t = 1$, we obtain:

$$\mathbb{E}[Q_1^*(s_1, \pi^*) - Q_1^*(s_1, \hat{\pi}_{\hat{\tau}})] \leq c \int \{Q_1^*(s, \pi^*) - Q_1^*(s, \hat{\pi}_{\hat{\tau}}\} \, ds.$$

Next we continue for generic t and bound the right hand side term of eqn. (10).

First we suppose we have a high-probability bound on ℓ_∞ convergence of $\hat{\tau}$. Define the good event

$$\mathcal{E}_g = \left\{ \sup_{s \in \mathcal{S}, a \in \mathcal{A}} |\hat{\tau}^{\hat{\underline{\pi}}_{t+1}}(s) - \tau^{\underline{\pi}^*_{t+1}, \circ}(s)| \leq K n^{-b*} \right\}$$

A maximal inequality gives that $P(\mathcal{E}_g) \geq 1 - n^{-\kappa}$. We have that

$$\int \{Q_t^*(s, \pi^*(s)) - Q_t^*(s, \hat{\pi}_{\hat{\tau}})\} \, ds = \int \{Q_t^*(s, \pi^*(s)) - Q_t^*(s, \hat{\pi}_{\hat{\tau}})\} \, \mathbb{I}\left[\mathcal{E}_g\right] ds + \int \{Q_t^*(s, \pi^*(s)) - Q_t^*(s, \hat{\pi}_{\hat{\tau}})\} \, \mathbb{I}\left[\mathcal{E}_g^c\right] ds \tag{11}$$

Assuming boundedness, the bad event occurs with vanishingly small probability $n^{-\kappa}$, which bounds the second term of eqn. (15).

For the first term of eqn. (15), note that on the good event, if mistakes occur such that $\pi_t^*(s) \neq \hat{\pi}_t(s)$, then the true contrast function is still bounded in magnitude by the good event ensuring closeness of the estimate, so that $\left|\tau_t^{\underline{\pi}^*_{t+1}, \circ}(s)\right| \leq 2Kn^{-b_*}$. And if no mistakes occur, at s the contribution to the integral is 0. Denote the mistake region as

$$\mathcal{S}_m = \{s \in \mathcal{S} \colon \left|\tau_t^{\underline{\pi}^*_{t+1}, \circ}(s)\right| \leq 2Kn^{-b_*}\}$$

Therefore

$$\int \{Q_t^*(s, \pi^*(s)) - Q_t^*(s, \hat{\pi}_{\hat{\tau}})\} \, ds \leq \int_{s \in \mathcal{S}_m} \{Q_t^*(s, \pi^*(s)) - Q_t^*(s, \hat{\pi}_{\hat{\tau}})\} \, \mathbb{I}\left[s \in \mathcal{S}_m\right] \mathbb{I}\left[\mathcal{E}_g\right] ds + O(n^{-\kappa}) \tag{12}$$

Note also that (for two actions), if *action* mistakes occur on the good event $\mathcal{E}_g$, the difference of Q functions must be near the decision boundaries so that we have the following bound on the integrand:

$$|Q^*(s, \pi^*) - Q^*(s, \hat{\pi})| \leq |\tau^{\underline{\pi}^*_{t+1}, \circ}| \leq 2Kn^{-b*}. \tag{13}$$

Therefore,

$$\begin{aligned} \int \{Q_t^*(s, \pi^*(s)) - Q_t^*(s, \hat{\pi}_{\hat{\tau}})\} \, ds &\leq O(n^{-\kappa}) + Kn^{-b*} \int \mathbb{I}\left[s \in \mathcal{S}_m\right] ds \\ &\leq O(n^{-\kappa}) + (Kn^{-b*})(Kn^{-b*\alpha}) \\ &= O(n^{-\kappa}) + (K^2 n^{-b*(1+\alpha)}) \end{aligned} \tag{14}$$

where the first inequality follows from the above, and the second from assumption 7 (margin).

Combining Eqs. (9) and (14), we obtain:

$$\begin{aligned} \mathbb{E}[V_t^*(S_t)] - \mathbb{E}[V_t^{\hat{\pi}_{\hat{\tau}}}(S_t)] &\leq \sum_{t=1}^{T} \gamma^t c \left\{ \int Q_t^{\hat{\pi}_{\hat{\tau}}}(s, \pi^*(s)) - Q_t^{\hat{\pi}_{\hat{\tau}}}(s, \hat{\pi}_{\hat{\tau}}) ds \right\} \\ &\leq \frac{(1 - \gamma^T)}{1 - \gamma} cT \{O(n^{-\kappa}) + (K^2 n^{-b*(1+\alpha)})\} \end{aligned}$$

We also obtain analogous results for norm bounds:

$$\begin{aligned} &\left\{ \int \left(Q_t^*(s, \pi^*(s)) - Q_t^*(s, \hat{\pi}_{\hat{\tau}})\right)^u ds \right\}^{1/u} \\ &\quad \leq \left\{ \int_{s \in \mathcal{S}_m} (Q_t^*(s, \pi^*(s)) - Q_t^*(s, \hat{\pi}_{\hat{\tau}}))^u \mathbb{I}\left[s \in \mathcal{S}_m\right] \mathbb{I}\left[\mathcal{E}_g\right] ds \right\}^{1/u} + O(n^{-\kappa}) \\ &\quad \leq \frac{(1 - \gamma^T)}{1 - \gamma} cT \{O(n^{-\kappa}) + (K^2 n^{-b*(1+\alpha)})\} \end{aligned}$$

So far we have made the somewhat stronger Assumption 9 (bounded transition density). Now we assume integrated risk convergence and the weaker sup-norm concetrability bound of Assumption 3. See (Munos and Szepesvári, 2008) for more discussion on the relationship between them.

The results under an integrated risk bound assumption on convergence of τ follow analogously as (Shi et al., 2022), which we also include for completeness. For a given $\varepsilon > 0$, redefine the mistake region parametrized by ϵ:

$$\mathcal{S}_\epsilon = \left\{ \max_a Q^*(s,a) - Q^*\left(s, \hat{\pi}(s)\right) \leq \varepsilon \right\}.$$

Again we obtain the bound by conditioning on the mistake region, and the triangle inequality:

$$\left\|Q_t^*\left(S_t, \pi^*\right) - Q_t^*\left(S_t, \hat{\pi}_{\hat{\tau}}\right)\right\|_2 \leq \left\|(Q_t^*\left(S_t, \pi^*\right) - Q_t^*\left(S_t, \hat{\pi}_{\hat{\tau}}\right))\mathbb{I}\left[\mathcal{S}_\epsilon\right]\right\|_2 + \left\|(Q_t^*\left(S_t, \pi^*\right) - Q_t^*\left(S_t, \hat{\pi}_{\hat{\tau}}\right))\mathbb{I}\left[\mathcal{S}_\epsilon^c\right]\right\|_2 \quad (15)$$

Using similar arguments as earlier, we can show by Assumption 7:

$$\left\|(Q_t^*\left(S_t, \pi^*\right) - Q_t^*\left(S_t, \hat{\pi}_{\hat{\tau}}\right))\mathbb{I}\left[\mathcal{S}_\epsilon\right]\right\|_2 \leq \left\|(Q_t^*\left(S_t, \pi^*\right) - Q_t^*\left(S_t, \hat{\pi}_{\hat{\tau}}\right))\mathbb{I}\left[\mathcal{S}_\epsilon\right]\right\|_1 \leq \varepsilon \mathbb{E}_{\pi_t^b}[\mathbb{I}[S_{t+1} \in \mathcal{S}_*]] = O\left(\varepsilon^{1+\alpha}\right).$$

As previously argued, we can show mistakes $\pi_t^*(s) \neq \hat{\pi}_t(s)$ occur only when

$$\max_a Q_t^*(s,a) - Q^*(s, \hat{\pi}(s)) \leq 2\left|\hat{\tau}^{\hat{\pi}_{t+1}}(s) - \tau^{\pi_{t+1}^*,\circ}(s)\right|. \quad (16)$$

It follows that

$$\begin{aligned}
\left\|(Q_t^*\left(S_t, \pi^*\right) - Q_t^*\left(S_t, \hat{\pi}_{\hat{\tau}}\right))\mathbb{I}\left[\mathcal{S}_\epsilon^c\right]\right\|_2 &\leq \left\|(Q_t^*\left(S_t, \pi^*\right) - Q_t^*\left(S_t, \hat{\pi}_{\hat{\tau}}\right))\mathbb{I}\left[\mathcal{S}_\epsilon^c\right]\right\|_1 \\
&\leq \mathbb{E}_{\pi_t^b}\left[\frac{4|\hat{\tau}^{\hat{\pi}_{t+1}}(s) - \tau^{\pi_{t+1}^*,\circ}(s)|^2}{|Q_t^*(s,\pi^*(s)) - Q_t^*(s,\hat{\pi}_{\hat{\tau}})|}\mathbb{I}\left[s \in \mathcal{S}_\epsilon^c\right]\right] \\
&\leq \frac{4}{\varepsilon}\mathbb{E}[|\hat{\tau}^{\hat{\pi}_{t+1}}(s) - \tau^{\pi_{t+1}^*,\circ}(s)|^2] = O\left(\varepsilon^{-1}|\mathcal{I}|^{-2b_*}\right).
\end{aligned}$$

Combining this yields that

$$\left\|Q_t^*\left(S_t, \pi^*\right) - Q_t^*\left(S_t, \hat{\pi}_{\hat{\tau}}\right)\right\|_2 \lesssim \varepsilon^{1+\alpha} + \varepsilon^{-1}|\mathcal{I}|^{-2b_*}$$

The result follows by choosing $\varepsilon = n^{-2b_*/(2+\alpha)}$ to balance the two terms.

For the norm bound, the first term is analogously bounded as $O\left(\varepsilon^{1+\alpha}\right)$:

$$\left\|(Q_t^*\left(S_t, \pi^*\right) - Q_t^*\left(S_t, \hat{\pi}_{\hat{\tau}}\right))\mathbb{I}\left[\mathcal{S}_\epsilon\right]\right\|_2 = O\left(\varepsilon^{1+\alpha}\right).$$

For the second term,

$$\begin{aligned}
\left\|(Q_t^*\left(S_t, \pi^*\right) - Q_t^*\left(S_t, \hat{\pi}_{\hat{\tau}}\right))\mathbb{I}\left[\mathcal{S}_\epsilon^c\right]\right\|_2 &\leq \left\{\mathbb{E}\left[\left(\frac{4|\hat{\tau}^{\hat{\pi}_{t+1}}(s) - \tau^{\pi_{t+1}^*,\circ}(s)|^2}{Q_t^*(s,\pi^*(s)) - Q_t^*(s,\hat{\pi}_{\hat{\tau}})}\right)^2 \mathbb{I}\left[s \in \mathcal{S}_\epsilon^c\right]\right]\right\}^{1/2} \\
&\leq \frac{4}{\varepsilon}\{\mathbb{E}[|\hat{\tau}^{\hat{\pi}_{t+1}}(s) - \tau^{\pi_{t+1}^*,\circ}(s)|^4]\}^{1/2} = O\left(\varepsilon^{-1}|\mathcal{I}|^{-2b_*}\right).
\end{aligned}$$

The result follows as previous by applying Assumption 3 to the sum decomposition of eqn. (9). □

Proof of Theorem 4.1. In the following, at times we omit the fixed evaluation policy π^e from the notation for brevity. That is, in this proof, $\hat{\tau}_t, \tau_t^n$ are equivalent to $\hat{\tau}_t^{\pi^e}, \tau_t^{n,\pi^e}$. Further define

$$\nu_t = \hat{\tau}_t - \tau_t^n, \nu_t^\circ = \hat{\tau}_t - \tau_t^\circ$$

Strong convexity of the squared loss implies that:

$$D_{\tau_t,\tau_t}\mathcal{L}\left(\tau_t, \hat{\eta}\right)[\nu_t, \nu_t] \geq \lambda \left\|\nu_t\right\|_2^2$$

therefore

$$\begin{aligned}\frac{\lambda}{2}\left\|\nu_t\right\|_2^2 &\leq \mathcal{L}_D(\hat{\tau}_t,\hat{\eta}) - \mathcal{L}_D(\tau_t^n,\hat{\eta}) - D_{\tau_t}\mathcal{L}_D(\tau_t^n,\hat{\eta})[\nu_t] \\ &\leq \epsilon(\hat{\tau}_t,\hat{\eta}) - D_{\tau_t}\mathcal{L}_D(\tau_t^n,\eta^\circ)[\nu_t] \\ &\quad + D_{\tau_t}\mathcal{L}_D(\tau_t^n,\eta^\circ)[\nu_t] - D_{\tau_t}\mathcal{L}_D(\tau_t^n,\hat{\eta})[\nu_t]\end{aligned} \tag{17}$$

We bound each term in turn.

To bound $\left|D_{\tau_t}\mathcal{L}_D(\tau_t^n,\eta^\circ)[\nu_t]\right|$, note that

$$D_{\tau_t}\mathcal{L}_D(\tau_t^n,\eta^\circ)[\nu_t] = \mathbb{E}[(R+\gamma Q_{t+1} - V_t^{\pi^b,\pi_{t+1:T}} + \left(A-\pi_t^b\right)\tau_t))\left(A-\pi_t^b\right)\nu_t]$$

and by the properties of the conditional moment at the true τ°,

$$= \mathbb{E}[(R+\gamma Q_{t+1} - V_t^{\pi^b,\pi_{t+1:T}} + \left(A-\pi_t^b\right)\tau_t^\circ))\left(A-\pi_t^b\right)\nu_t] = 0$$

Therefore,

$$D_{\tau_t}\mathcal{L}_D(\tau_t^n,\eta^\circ)[\nu_t] = -\mathbb{E}[(\tau^\circ-\tau_t^n)(A-\pi_t^b)(A-\pi_t^b)(\hat{\tau}_t-\tau_t^n)]$$

Note that in general, for generic p,q,r such that $1/p+1/q+1/r=1$ we have that $\mathbb{E}[fgh] \leq \|fg\|_{p'}\|h\|_r \leq \|f\|_p\|g\|_q\|h\|_r$ where $p' = \frac{pq}{p+q}$ or $\frac{1}{p'} = \frac{1}{p}+\frac{1}{q}$ or $1 = \frac{1}{p/p'}+\frac{1}{q/p'}$.

Therefore,

$$\begin{aligned}D_{\tau_t}\mathcal{L}_D(\tau_t^n,\eta^\circ)[\nu_t] &\leq \left|D_{\tau_t}\mathcal{L}_D(\tau_t^n,\eta^\circ)[\nu_t]\right| \\ &\leq \mathbb{E}[(\tau^\circ-\tau_t^n)\mathbb{E}[(A_t-\pi_t^b)(A_t-\pi_t^b)\mid S_t](\hat{\tau}_t-\tau_t^n)] \\ &\leq \|(\tau^\circ-\tau_t^n)\|_u\|(\hat{\tau}_t-\tau_t^n)\|_{\overline{u}}\cdot\left\{\sup_s \mathbb{E}[(A_t-\pi_t^b)(A_t-\pi_t^b)\mid s]\right\}\end{aligned}$$

where $u,\overline{u}$ satisfy $\frac{1}{u}+\frac{1}{\overline{u}}=1$.

Next we bound $D_{\tau_t}\mathcal{L}_D(\tau_t^n,\eta^\circ)[\nu_t] - D_{\tau_t}\mathcal{L}_D(\tau_t^n,\hat{\eta})[\nu_t]$ by universal orthogonality. By a second order Taylor expansion, we have that, where $\eta_\epsilon = \eta^\circ + \epsilon(\hat{\eta}-\eta^\circ)$.

$$D_{\tau_t}\left(\mathcal{L}_D(\tau_t^n,\eta^\circ) - \mathcal{L}_D(\tau_t^n,\hat{\eta})\right)[\nu_t] = \frac{1}{2}\int_0^1 D_{\eta,\eta,\tau_t}(\tau_t^n,\tau_{t+1}^\circ,\eta_\epsilon)[\hat{\eta}-\eta^\circ,\hat{\eta}-\eta^\circ,\nu_t]$$

We can deduce from Lemmas 4 and 5 that the integrand is:

$$\begin{aligned}&\mathbb{E}\left[\tau_t^2\left(\hat{\pi}_t^b-\pi_t^{b,\circ}\right)^2\nu_t\right] + \mathbb{E}\left[(\hat{\pi}_t^b-\pi_t^{b,\circ})\tau_t(\hat{m}_t-m_t^\circ)\nu_t\right] + \mathbb{E}\left[(\hat{\pi}_t^b-\pi_t^{b,\circ})\tau_t\gamma(\hat{Q}_{t+1}-Q_{t+1}^\circ)\nu_t\right] \\ &\quad - \mathbb{E}\left[(\hat{m}_t-m_t^\circ)\gamma\left(\hat{Q}_{t+1}-Q_{t+1}^\circ\right)\nu_t\right] \\ &\leq {B_\tau}^2\|\left(\hat{\pi}_t^b-\pi_t^{b,\circ}\right)^2\|_u\|\nu_t\|_{\overline{u}} + B_\tau\|(\hat{\pi}_t^b-\pi_t^{b,\circ})(\hat{m}_t-m_t^\circ)\|_u\|\nu_t\|_{\overline{u}} + \gamma B_\tau\|(\hat{\pi}_t^b-\pi_t^{b,\circ})(\hat{Q}_{t+1}-Q_{t+1}^\circ)\|_u\|\nu_t\|_{\overline{u}} \\ &\quad + \gamma\|(\hat{m}_t-m_t^\circ)(\hat{Q}_{t+1}-Q_{t+1}^\circ)\|_u\|\nu_t\|_{\overline{u}}\end{aligned}$$

Putting the bounds together, we obtain:

$$\begin{aligned}\frac{\lambda}{2}\left\|\nu_t\right\|_2^2 &\leq \epsilon(\hat{\tau}_t,\hat{\eta}) + \|\nu_t\|_{\overline{u}}\|(\tau^\circ-\tau_t^n)\|_u \\ &\quad + \|\nu_t\|_{\overline{u}}\Bigg({B_\tau}^2\|\left(\hat{\pi}_t^b-\pi_t^{b,\circ}\right)^2\|_u + B_\tau\|(\hat{\pi}_t^b-\pi_t^{b,\circ})(\hat{m}_t-m_t^\circ)\|_u + \gamma B_\tau\|(\hat{\pi}_t^b-\pi_t^{b,\circ})(\hat{Q}_{t+1}-Q_{t+1}^\circ)\|_u \\ &\quad + \gamma\|(\hat{m}_t-m_t^\circ)(\hat{Q}_{t+1}-Q_{t+1}^\circ)\|_u\Bigg)\end{aligned} \tag{18}$$

Let $\rho_t^{\pi^e}(\hat{\eta})$ denote the collected product error terms, e.g.

$$\rho_t^{\pi^e}(\hat{\eta}) = B_\tau{}^2 \|\left(\hat{\pi}_t^b - \pi_t^{b,\circ}\right)^2\|_u + B_\tau \|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{m}_t - m_t^\circ)\|_u$$
$$+ \gamma(B_\tau \|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{Q}_{t+1} - Q_{t+1}^\circ)\|_u + \|(\hat{m}_t - m_t^\circ)(\hat{Q}_{t+1} - Q_{t+1}^\circ)\|_u)$$

Analogously we drop the π^e decoration from ρ_t in this proof. The AM-GM inequality implies that for $x, y \geq 0$, $\sigma > 0$, we have that $xy \leq \frac{1}{2}(\frac{2}{\sigma}x^2 + \frac{\sigma}{2}y^2)$. Therefore

$$\frac{\lambda}{2}\|\nu_t\|_2^2 - \frac{\sigma}{4}\|\nu_t\|_{\bar{u}}^2 \leq \epsilon(\hat{\tau}_t, \hat{\eta}) + \frac{1}{\sigma}\left(\|(\tau^\circ - \tau_t^n)\|_u + \rho_t(\hat{\eta})\right)^2 \tag{19}$$

and since $(x+y)^2 \leq 2(x^2+y^2)$,

$$\frac{\lambda}{2}\|\nu_t\|_2^2 - \frac{\sigma}{4}\|\nu_t\|_{\bar{u}}^2 \leq \epsilon(\hat{\tau}_t, \hat{\eta}) + \frac{2}{\sigma}\left(\|(\tau^\circ - \tau_t^n)\|_u^2 + \rho_t(\hat{\eta})^2\right)$$

□

Proof of Theorem B.2. Let $\hat{\mathcal{L}}_{S,t}, \hat{\mathcal{L}}_{S',t}$ denote the empirical loss over the samples in S and S'; analogously $\hat{\eta}_S, \hat{\eta}_{S'}$ are the nuisance functions trained on each sample split.

Define the loss function ℓ_t on observation $O = \{(S_t, A_t, R_t, S_{t+1})\}_{t=1}^T$:

$$\ell_t(O; \tau_t; \hat{\eta}) = \left(\{R_t + \hat{Q}_{t+1}^{\pi_{t+1}^e}(S_{t+1}, A_{t+1}) - \hat{m}_t(S_t)\} - \{A - \hat{\pi}_t^b(1 \mid S_t)\} \cdot \tau_t(S_t)\right)^2$$

and the centered loss function $\Delta\ell$, centered with respect to $\hat{\tau}_t^n$:

$$\Delta\ell_t(O; \tau_t; \hat{\eta}) = \ell_t(O; \tau_t; \hat{\eta}) - \ell_t(O; \hat{\tau}_t^n; \hat{\eta}).$$

Assuming boundedness, ℓ_t is $L-$Lipschitz constant in τ_t:

$$|\Delta\ell_t(O; \tau_t; \hat{\eta}) - \Delta\ell_t(O; \tau_t'; \hat{\eta})| \leq L\|\tau_t - \tau_t\|_2.$$

Note that $\ell(O, \hat{\tau}_t^n, \hat{\eta}) = 0$. Define the centered average losses:

$$\Delta\hat{\mathcal{L}}_{S,t}(\tau_t, \hat{\eta}) = \hat{\mathcal{L}}_{S,t}(\tau_t, \hat{\eta}) - \hat{\mathcal{L}}_{S,t}(\hat{\tau}_t^n, \hat{\eta}) = \hat{\mathbb{E}}_{n/2}^S[\Delta\ell_t(O, \tau_T, \hat{\eta})]$$
$$\Delta\mathcal{L}_{S,t}(\tau_t, \hat{\eta}) = \mathcal{L}_{S,t}(\tau_t, \hat{\eta}) - \mathcal{L}_{S,t}(\hat{\tau}_t^n, \hat{\eta}) = \mathbb{E}[\Delta\ell_t(O, \tau_T, \hat{\eta})]$$

Assume that δ_n is an upper bound on the critical radius of the centered function class $\{\Psi_{t,i}^n - \hat{\tau}_{t,i}^n$, with $\delta_n = \Omega(\frac{r\log\log n}{n})$, and define $\delta_{n,\xi} = \delta_n + c_0\sqrt{\frac{\log(c_1 T/\xi)}{n}}$ for some c_0, c_1.

By Lemma 7 (Lemma 14 of (Foster and Syrgkanis, 2019) on local Rademacher complexity decompositions), with high probability 1-ξ, for all $t \in [T]$, and for c_0 a universal constant ≥ 1.

$$|\Delta\mathcal{L}_{S,t}(\hat{\tau}_t, \hat{\eta}_{S'}) - \Delta\mathcal{L}_{D,t}(\hat{\tau}_t, \hat{\eta}_{S'})| = |\Delta\mathcal{L}_{S,t}(\hat{\tau}_t, \hat{\eta}_{S'}) - \Delta\mathcal{L}_{S,t}(\hat{\tau}_t^n, \hat{\eta}_{S'}) - (\Delta\mathcal{L}_{D,t}(\hat{\tau}_t, \hat{\eta}_{S'}) - \Delta\mathcal{L}_{D,t}(\hat{\tau}_t^n, \hat{\eta}_{S'}))|$$
$$\leq c_0\left(rm\delta_{n/2,\xi}\|\hat{\tau}_t - \hat{\tau}_t^n\|_2^2 + rm\delta_{n/2,\xi}^2\right)$$

Assuming realizability of $\hat{\tau}_t$, we have that $\frac{1}{2}\left(\Delta\hat{\mathcal{L}}_{S,t}(\hat{\tau}_t, \hat{\eta}_{S'}) + \Delta\hat{\mathcal{L}}_{S',t}(\hat{\tau}_t, \hat{\eta}_S)\right) \leq 0$. Then with high probability $\geq 1 - 2\xi$:

$$\frac{1}{2}\left(\Delta\mathcal{L}_{D,t}(\hat{\tau}_t, \hat{\eta}_{S'}) + \Delta\mathcal{L}_{D,t}(\hat{\tau}_t, \hat{\eta}_S)\right)$$
$$\leq \frac{1}{2}|\Delta\mathcal{L}_{D,t}(\hat{\tau}_t, \hat{\eta}_{S'}) - \Delta\mathcal{L}_{S,t}(\hat{\tau}_t, \hat{\eta}_{S'}) + \Delta\mathcal{L}_{D,t}(\hat{\tau}_t, \hat{\eta}_S) - \Delta\mathcal{L}_{S',t}(\hat{\tau}_t, \hat{\eta}_S)|$$
$$\leq \frac{1}{2}|\Delta\mathcal{L}_{D,t}(\hat{\tau}_t, \hat{\eta}_{S'}) - \Delta\mathcal{L}_{S,t}(\hat{\tau}_t, \hat{\eta}_{S'})| + |\Delta\mathcal{L}_{D,t}(\hat{\tau}_t, \hat{\eta}_S) - \Delta\mathcal{L}_{S',t}(\hat{\tau}_t, \hat{\eta}_S)|$$
$$\leq c_0\left(rm\delta_{n/2,\xi}\|\hat{\tau}_t - \hat{\tau}_t^n\|_2 + rm\delta_{n/2,\xi}^2\right)$$

The ϵ excess risk term in Theorem 4.1 indeed corresponds to one of the loss differences defined here, i.e. $\Delta\mathcal{L}_{D,t}(\hat{\tau}_t, \hat{\eta}_S) := \epsilon(\hat{\tau}_t^n, \hat{\tau}_t, \hat{h}_S)$. Therefore, applying Theorem 4.1 with $u = \overline{u} = 2$ and $\sigma = \lambda$ with the above bound, and averaging the sample-split estimators, we obtain

$$\frac{\lambda}{4}\|\nu_t\|_2^2 \leq \frac{1}{2}\left(\epsilon(\hat{\tau}_t, \hat{\eta}_S) + \epsilon(\hat{\tau}_t, \hat{\eta}_{S'})\right) + \frac{2}{\lambda}\left(\|\tau_t^\circ - \hat{\tau}_t^n\|_2^2 + \sum_{s\in\{S,S'\}} \rho_t(\hat{\eta}_s)^2\right)$$

We further decompose the excess risk of empirically-optimal $\hat{\tau}_t$ relative to the population minimizer to instead bound by the error of $\hat{\tau}_t$ to the projection onto Ψ, $\hat{\tau}_t^n$, since $\|\hat{\tau}_t - \tau_t^\circ\|_2^2 \leq \|\hat{\tau}_t - \hat{\tau}_t^n\|_2^2 + \|\hat{\tau}_t^n - \tau_t^\circ\|_2^2$, we obtain

$$\frac{\lambda}{4}\|\hat{\tau}_t - \tau_t^\circ\|_2^2 \leq c_0\left(rm\delta_{n/2,\xi}\|\hat{\tau}_t - \hat{\tau}_t^n\|_2 + rm\delta_{n/2,\xi}^2\right) + \frac{8+\lambda^2}{4\lambda}\|\tau_t^\circ - \tau_t^n\|_2^2 + \frac{2}{\lambda}\sum_{s\in\{S,S'\}}\rho_t(\hat{\eta}_s)^2$$

Again using the AM-GM inequality $xy \leq \frac{1}{2}\left(\frac{2}{\sigma}x^2 + \frac{\sigma}{2}y^2\right)$, we bound

$$\begin{aligned} c_0\left(rm\delta_{n/2,\xi}\|\hat{\tau}_t - \hat{\tau}_t^n\|_2 + rm\delta_{n/2,\xi}^2\right) &\leq \frac{c_0}{2}r^2m^2(1+\frac{2}{\epsilon})\delta_{n/2,\xi}^2 + \frac{\epsilon}{4}\|\hat{\tau}_t - \hat{\tau}_t^n\|_2^2 \\ &\leq c_0r^2m^2(1+\frac{1}{\epsilon})\delta_{n/2,\xi}^2 + \frac{\epsilon}{4}(\|\hat{\tau}_t - \tau_t^\circ\|_2^2 + \|\tau_t^\circ - \hat{\tau}_t^n\|_2^2) \end{aligned}$$

Therefore,

$$\frac{\lambda-\epsilon}{4}\|\hat{\tau}_t - \tau_t^\circ\|_2^2 \leq c_0r^2m^2(1+\frac{1}{\epsilon})\delta_{n/2,\xi}^2 + \left(\frac{8+\lambda^2}{4\lambda} + \frac{\epsilon}{4}\right)\|\tau_t^\circ - \tau_t^n\|_2^2 + \frac{2}{\lambda}\sum_{s\in\{S,S'\}}\rho_t(\hat{\eta}_s)^2$$

Choose $\epsilon \leq \lambda/8$ so that

$$\begin{aligned} \frac{\lambda}{8}\|\hat{\tau}_t - \tau_t^\circ\|_2^2 &\leq c_0r^2m^2(1+\frac{8}{\lambda})\delta_{n/2,\xi}^2 + \left(\frac{4+\lambda^2}{2\lambda}\right)\|\tau_t^\circ - \tau_t^n\|_2^2 + \frac{2}{\lambda}\sum_{s\in\{S,S'\}}\rho_t(\hat{\eta}_s)^2 \\ &\leq \left(1+\frac{8}{\lambda}+\frac{\lambda}{2}\right)(c_0r^2m^2\delta_{n/2,\xi}^2 + \|\tau_t^\circ - \tau_t^n\|_2^2 + \sum_{s\in\{S,S'\}}\rho_t(\hat{\eta}_s)^2) \end{aligned}$$

and therefore

$$\|\hat{\tau}_t - \tau_t^\circ\|_2^2 \leq \left(\frac{8}{\lambda}(1+\frac{8}{\lambda}) + 4\right)(c_0r^2m^2\delta_{n/2,\xi}^2 + \|\tau_t^\circ - \tau_t^n\|_2^2 + \sum_{s\in\{S,S'\}}\rho_t(\hat{\eta}_s)^2)$$

Taking expectations:

$$\mathbb{E}[\|\hat{\tau}_t - \tau_t^\circ\|_2^2] \leq \left(\frac{8}{\lambda}(1+\frac{8}{\lambda}) + 4\right)(c_0r^2m^2\delta_{n/2}^2 + \|\tau_t^\circ - \tau_t^n\|_2^2 + \max_{s\in\{S,S'\}}\mathbb{E}[\rho_t(\hat{\eta}_s)^2])$$

Therefore, if the product error rate terms are all of the same order as the estimation order terms:

$$\begin{aligned} &\mathbb{E}[\|\hat{\pi}_t^b - \pi_t^{b,\circ}\|_2^2] = O(\delta_{n/2}^2 + \|\tau_t^\circ - \tau_t^n\|_2^2) \\ &\mathbb{E}[\|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{m}_t - m_t^\circ)\|_2^2] = O(\delta_{n/2}^2 + \|\tau_t^\circ - \tau_t^n\|_2^2) \\ &\mathbb{E}[\|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{Q}_{t+1} - Q_{t+1}^\circ)\|_2^2] = O(\delta_{n/2}^2 + \|\tau_t^\circ - \tau_t^n\|_2^2) \\ &\mathbb{E}[\|(\hat{m}_t - m_t^\circ)(\hat{Q}_{t+1} - Q_{t+1}^\circ)\|_2^2] = O(\delta_{n/2}^2 + \|\tau_t^\circ - \tau_t^n\|_2^2) \end{aligned}$$

□

Proof of Theorem 4.2. **Preliminaries** We introduce some additional notation. For the analysis of implications of policy optimization, we further introduce notation that parametrizes the time-t loss function with respect to

the time-$(t+1)$ policy. In analyzing the policy optimization, this will be used to decompose the policy error arising from time steps closer to the horizon. Define

$$\mathcal{L}_D(\tau_t^n, \tau'_{t+1}, \hat{\eta}) = \mathbb{E}\left[\left(\{R_t + \gamma Q_{t+1}^{\pi_{\tau'_{t+1}}}(S_{t+1}, A_{t+1}) - V_{\pi_t^b, \pi_{\tau'_{t+1}}}(S_t)\} - \{A - \pi_t^b(1 \mid S_t)\} \cdot \tau(S_t)\right)^2\right]$$

where $\pi_{\tau'_{t+1}}(s) \in \operatorname{argmax} \tau'_{t+1}(s)$. That is, the second argument parameterizes the difference-of-Q function that generates the policy that oracle nuisance functions are evaluated at.

Then, for example, the true optimal policy satisfies that $\pi_t^* \in \arg\max \tau_t^\circ(s)$. We define the oracle loss function with nuisance functions evaluated with respect to the optimal policy π^*.

$$\mathcal{L}_D(\tau_t^n, \tau^\circ, \hat{\eta}) = \mathbb{E}\left[\left(\{R_t + \gamma Q_{t+1}^{\pi^*_{\tau^\circ_{t+1}}}(S_{t+1}, A_{t+1}) - m^\circ(S_t)\} - \gamma\{A - \pi_t^b(1 \mid S_t)\} \cdot \tau(S_t)\right)^2\right]$$

In contrast, the empirical policy optimizes with respect to a next-stage *estimate* of the *empirical best* next-stage policy $\hat{\pi}_{\hat{\tau}_{t+1}}$. That is, noting the empirical loss function:

$$\mathcal{L}_D(\tau_t^n, \hat{\tau}_{t+1}, \hat{\eta}) = \mathbb{E}\left[\left(\{R_t + \gamma Q_{t+1}^{\hat{\pi}_{\hat{\tau}_{t+1}}}(S_{t+1}, A_{t+1}) - m^\circ(S_t)\} - \gamma\{A - \pi_t^b(1 \mid S_t)\} \cdot \tau(S_t)\right)^2\right]$$

Step 1: Applying advantage estimation results. At every timestep, the first substep is to estimate the Q-function contrast, $\hat{\tau}_t^{\hat{\pi}_{t+1}}$. The assumptions on product error nuisance rates imply that for a fixed $\hat{\pi}_{t+1}$ that we would obtain estimation error

$$\mathbb{E}\left[\|\hat{\tau}_t^{\hat{\pi}_{t+1}} - \tau_t^{\hat{\pi}_{t+1},\circ}\|_2^2\right] = O\left(\delta_{n/2}^2 + \left\|\tau_t^{\pi^e,\circ} - \tau_t^{\pi^e,n}\right\|_2^2\right)$$

Step 2: Establishing policy consistency. Applying Lemma 1 requires a convergence rate of $\hat{\tau}_t^{\hat{\pi}_{t+1}}$ to $\hat{\tau}_t^{\pi^*_{t+1}}$. The estimation error guarantees on the contrast function, however, are for the policy $\hat{\pi}_{t+1}$. We obtain the required bound via induction. At a high level, the estimation error arising from $\hat{\pi}_{t+1}$ vs π^*_{t+1} too eventually is integrated; so when the margin exponent $\alpha > 0$, these policy error terms are higher-order and vanish at a faster rate.

Importantly, we suppose the product error rate conditions hold for each t for data-optimal policies evaluated along the algorithm, i.e. for each t, for each t, for $\underline{\hat{\pi}}_{t+1}$, each of $\mathbb{E}[\|(\hat{\pi}_t^b - \pi_t^{b,\circ})\|_2^2]$, $\mathbb{E}[\|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{m}_t^{\underline{\hat{\pi}}_{t+1}} - m_t^{\circ,\underline{\hat{\pi}}_{t+1}})\|_2^2]$, $\mathbb{E}[\|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{Q}_{t+1}^{\underline{\hat{\pi}}_{t+2}} - Q_{t+1}^{\circ,\underline{\hat{\pi}}_{t+2}})\|_2^2]$, and $\mathbb{E}[\|(\hat{m}_t - m_t^\circ)(\hat{Q}_{t+1}^{\circ,\underline{\hat{\pi}}_{t+2}} - Q_{t+1}^{\circ,\underline{\hat{\pi}}_{t+2}})\|_2^2]$ are of order $O(\delta_{n/2}^2 + \|\tau_t^{\hat{\pi}_{t+1},\circ} - \tau_t^{\hat{\pi}_{t+1},n}\|_2^2)$.

Step 2a: induction hypothesis.

Next we show the induction hypothesis.

First we consider the base case: When $t = T$, τ_T is independent of the forward policy so that $\|\hat{\tau}_T^{\hat{\pi}} - \tau_T^{\circ,\pi^*}\| = \|\hat{\tau}_T - \tau_T^\circ\|$. Then the base case follows by Theorem B.2.

Suppose it is true that for timesteps $k \geq t+1$, we have that

$$\|\hat{\tau}_k^{\underline{\hat{\pi}}_{k+1}} - \tau_k^{\circ,\underline{\pi}^*_{k+1}}\| = O(\delta_{n/2} + \|\tau_k^{\circ,\underline{\hat{\pi}}_{k+1}} - \tau_k^{n,\underline{\hat{\pi}}_{k+1}}\|_2) + Kn^{-\mathcal{R}_k}, \tag{20}$$

where

$$\mathcal{R}_k = \min\left(\rho_{k+1}^{(c)} \cdot \frac{2+2\alpha}{2+\alpha},\ \rho_{k+1}^{(\Psi)} \cdot \frac{2+2\alpha}{2+\alpha},\ -\{\min_{k' \geq k+1}(\rho_{k'}^{(c)}, \rho_{k'}^{(\Psi)})\} \cdot \frac{2+2\alpha}{2+\alpha}^{T-k'}\right). \tag{21}$$

And therefore, applying Lemma 1, that

$$\left|\mathbb{E}[V_k^{\pi^*} - V_k^{\hat{\pi}_{\hat{\tau}}}]\right| = O(n^{-\min\{\rho_k^{(c)}, \rho_k^{(\Psi)}\}\frac{2+2\alpha}{2+\alpha}}) + o(n^{-\min\{\rho_k^{(c)}, \rho_k^{(\Psi)}\}\frac{2+2\alpha}{2+\alpha}}). \tag{22}$$

We will show that the induction hypothesis implies

$$\|\hat{\tau}_t^{\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\pi}^*_{t+1}}\| \leq O(\delta_{n/2} + \|\tau_t^{\circ,\underline{\hat{\pi}}_{t+1}} - \tau_t^{n,\underline{\hat{\pi}}_{t+1}}\|_2) + Kn^{-\mathcal{R}_t}.$$

and

$$\left|\mathbb{E}[V_k^{\pi^*} - V_k^{\hat{\pi}_{\hat{\tau}}}]\right| = O(n^{-\min\{\rho_k^{(c)},\rho_k^{(\Psi)}\}\frac{2+2\alpha}{2+\alpha}}) + o(n^{-\min\{\rho_k^{(c)},\rho_k^{(\Psi)}\}\frac{2+2\alpha}{2+\alpha}})$$

First decompose the desired error $\|\hat{\tau}_t^{\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\pi}^*_{t+1}}\|$ as:

$$\|\hat{\tau}_t^{\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\pi}^*_{t+1}}\| \leq \|\hat{\tau}_t^{\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\hat{\pi}}_{t+1}}\| + \|\tau_t^{\circ,\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\pi}^*_{t+1}}\| \tag{23}$$

The first term is the policy evaluation estimation error, and under the product error rate assumptions , Theorems 4.1 and B.2 give that $\mathbb{E}[\|\hat{\tau}_t^{\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\hat{\pi}}_{t+1}}\|_2^2] = O(\delta_{n/2}^2 + \|\tau_t^{\circ,\underline{\hat{\pi}}_{t+1}} - \tau_t^{n,\underline{\hat{\pi}}_{t+1}}\|_2^2)$. The second term of the above depends on the convergence of the empirically optimal policy $\hat{\pi}$; we use our analysis from Lemma 1 to bound the impact of future estimates of difference-of-Q functions using the induction hypothesis. The following analysis will essentially reveal that the margin assumption of Assumption 7 implies that the error due to the empirically optimal policy is higher-order, and the first term (time$-t$ estimation error of $\hat{\tau}_t$) is the leading term.

As in eqn. (9), we have that:

$$V_t^*(s) - V_t^{\pi_{\hat{\tau}}}(s) \leq \gamma \mathbb{E}_{\hat{\pi}_t}\left[V_{t+1}^{\pi^*_1} - V_{t+1}^{\hat{\pi}_{\hat{\tau}}} \mid s_t\right] + Q_t^*(s,\pi^*) - Q_t^*(s,\hat{\pi}_{\hat{\tau}}).$$

Decompose:

$$\|\tau_t^{\circ,\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\pi}^*_{t+1}}\| \leq \sum_a \|Q_t^{\underline{\pi}^*_{t+1}}(s,a) - Q_t^{\underline{\hat{\pi}}_{t+1}}(s,a)\|$$

By definition of τ and $\pm V_{t+1}^{\hat{\pi}_{t+1},\underline{\pi}^*_{t+2}}$, for each a, we have that

$$\begin{aligned}
&\|Q_t^{\underline{\pi}^*_{t+1}}(s,a) - Q_t^{\underline{\hat{\pi}}_{t+1}}(s,a)\| \\
&= \|\mathbb{E}_{\pi_t^a}[V_{t+1}^{\underline{\pi}^*_{t+1}} - V_{t+1}^{\underline{\hat{\pi}}_{t+1}} \mid S_t]\| \\
&\leq \|\mathbb{E}_{\pi_t^a}[V_{t+1}^{\underline{\pi}^*_{t+1}} - V_{t+1}^{\hat{\pi}_{t+1},\underline{\pi}^*_{t+2}} \mid S_t]\| + \|\mathbb{E}_{\pi_t^a}[V_{t+1}^{\hat{\pi}_{t+1},\underline{\pi}^*_{t+2}} - V_{t+1}^{\underline{\hat{\pi}}_{t+1}} \mid S_t]\| \\
&= \|\mathbb{E}_{\pi_t^a}[Q_{t+1}^{\underline{\pi}^*_{t+2}}(S_{t+1},\pi^*_{t+1}) - Q_{t+1}^{\underline{\pi}^*_{t+2}}(S_{t+1},\hat{\pi}_{t+1}) \mid S_t]\| + \gamma\|\mathbb{E}_{\pi_t^a}[\mathbb{E}_{\underline{\hat{\pi}}_{t+1}}[V_{t+2}^{\underline{\pi}^*_{t+2}} - V_{t+2}^{\underline{\hat{\pi}}_{t+2}} \mid S_t]]\| && (24) \\
&\leq C_\infty \left\{\int (Q_{t+1}^{\underline{\pi}^*_{t+2}}(s,\pi^*_{t+1}) - Q_{t+1}^{\underline{\pi}^*_{t+2}}(s,\hat{\pi}_{t+1}))^2 ds\right\}^{1/2} + \gamma\|\mathbb{E}_{\pi_t^a}[\mathbb{E}_{\underline{\hat{\pi}}_{t+1}}[V_{t+2}^{\underline{\pi}^*_{t+2}} - V_{t+2}^{\underline{\hat{\pi}}_{t+2}} \mid S_t]]\| && (25)
\end{aligned}$$

where the last inequality follows by Assumption 9 and the policy-convolved transition density.

Next we bound the first term using the margin analysis of Lemma 1 and the inductive hypothesis. Supposing the product error rates are satisfied on the nuisance functions for estimation of $\hat{\tau}_{t+1}$, the induction hypothesis gives that

$$\mathbb{E}[\|\hat{\tau}_{t+1}^{\underline{\hat{\pi}}_{t+2}} - \tau_{t+1}^{\circ,\underline{\pi}^*_{t+2}}\|_2] = O\left(\delta_{n/2} + \|\tau_t^{\pi^e,\circ} - \tau_t^n\|_2 + n^{-\mathcal{R}_{t+1}}\right).$$

The induction hypothesis gives the integrated risk rate assumption on $\hat{\tau}_{t+1}$ to apply Lemma 1,

$$\begin{aligned}
&\left\{\int (Q_{t+1}^{\underline{\pi}^*_{t+2}}(s,\pi^*_{t+1}) - Q_{t+1}^{\underline{\pi}^*_{t+2}}(s,\hat{\pi}_{t+1}))^2 ds\right\}^{1/2} \\
&\quad \leq \frac{(1-\gamma^{T-t-1})}{1-\gamma} C_\infty (T-t-1)\{O(n^{-\kappa}) + Kn^{-\min\{r_{t+1}^{(c)},r_{t+1}^{(\Psi)},\mathcal{R}_{t+1}\}(1+\alpha)}\}.
\end{aligned}$$

Combining with the previous analysis, we obtain:

$$\begin{aligned}
\|\hat{\tau}_t^{\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\pi}^*_{t+1}}\|_2^2 \leq{}& O(\delta_{t,n/2}^2 + \|\tau_t^{\circ,\underline{\hat{\pi}}_{t+1}} - \tau_t^{n,\underline{\hat{\pi}}_{t+1}}\|_2^2) + O(n^{-\min\left\{\rho_{t+2}^{(c)},\rho_{t+2}^{(\Psi)},\mathcal{R}_{t+2}\right\}\frac{2+2\alpha}{2+\alpha}})\} \\
&+ \frac{(1-\gamma^{T-t-1})}{1-\gamma} C_\pi (T-t-1)\{O(n^{-\kappa}) + Kn^{-\min\{\rho_{t+1}^{(c)},\rho_{t+1}^{(\Psi)},\mathcal{R}_{t+1}\}\frac{2+2\alpha}{2+\alpha}}\}
\end{aligned} \tag{26}$$

from eqn. (24) and Appendix C.3.

Hence we obtain the inductive step and the result follows.

If we further assume that for $t' \geq t$, we have that $\rho_t^{(\cdot)} \leq \rho_{t'}^{(\cdot)}$, for $(\cdot) \in \{(c), (\Psi)\}$, i.e. the estimation error rate is nonincreasing over time, and that $\alpha > 0$ (i.e. Assumption 7, the margin assumption, holds with exponent $\alpha > 0$, then we can see from the result that the integrated risk terms obtain faster rates, hence are higher-order, and the leading term is the auxiliary estimation error of the Q-function contrast.

□

D Results used from other works

Here we collect technical lemmas from other works, stated without proof.

Lemma 6 (Lemma 18 of (Lewis and Syrgkanis, 2021))**.** *Consider any sequence of non-negative numbers $a_1, \ldots, a_m$ satisfying the inequality:*

$$a_t \leq \mu_t + c_t \max_{j=t+1}^{m} a_j$$

with $\mu_t, c_t \geq 0$. Let $c := \max_{t\in[m]} c_t$ and $\mu := \max_{t\in[m]} \mu_t$. Then it must also hold that:

$$a_t \leq \mu \frac{c^{m-t+1} - 1}{c - 1}$$

Lemma 7 (Lemma 14 of (Foster and Syrgkanis, 2019), see also results on local Rademacher complexity (Wainwright, 2019))**.** *Consider a function class $\mathcal{F}$, with $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq 1$, and pick any $f^\star \in \mathcal{F}$. Let $\delta_n^2 \geq \frac{4d\log(41\log(2c_2 n))}{c_2 n}$ be any solution to the inequalities:*

$$\forall t \in \{1, \ldots, d\} : \mathcal{R}\left(\text{star}\left(\mathcal{F}|_t - f_t^\star\right), \delta\right) \leq \delta^2.$$

Moreover, assume that the loss ℓ is L-Lipschitz in its first argument with respect to the ℓ_2 norm. Then for some universal constants c_5, c_6, with probability $1 - c_5 \exp\left(c_6 n \delta_n^2\right)$,

$$|\mathbb{P}_n\left(\mathcal{L}_f - \mathcal{L}_{f^\star}\right) - \mathbb{P}\left(\mathcal{L}_f - \mathcal{L}_{f^\star}\right)| \leq 18Ld\delta_n \left\{\|f - f^\star\|_2 + \delta_n\right\}, \quad \forall f \in \mathcal{F}.$$

Hence, the outcome $\hat{f}$ of constrained ERM satisfies that with the same probability,

$$\mathbb{P}\left(\mathcal{L}_{\hat{f}} - \mathcal{L}_{f^\star}\right) \leq 18Ld\delta_n \left\{\left\|\hat{f} - f^\star\right\|_2 + \delta_n\right\}.$$

If the loss $\mathcal{L}_f$ is also linear in f, i.e. $\mathcal{L}_{f+f'} = \mathcal{L}_f + \mathcal{L}_{f'}$ and $\mathcal{L}_{\alpha f} = \alpha \mathcal{L}_f$, then the lower bound on δ_n^2 is not required.

E Experimental details

All experiments were ran either on a Macbook Pro M1 with 16gb RAM and 8 CPU cores or on a computer cluster with 64 CPU cores of 8gb RAM each. Experiments were run in Python using native Python, CVXPY, and scikit-learn. Each figure took approximately 3-10 minutes to generate.

1d validation example (Appendix E) Following the specification of (Kallus and Uehara, 2019a, Sec 5.1), we consider a small MDP of $T = 30$, binary actions, univariate continuous state, initial state distribution $p(s_0) \sim \mathcal{N}(0.5, 0.2)$, transition probabilities $P_t(s_{t+1} \mid s_t, a_t) \sim \mathcal{N}(s + 0.3a - 0.15, 0.2)$. The target and behavior policies we consider are $\pi^e(a \mid s) \sim \text{Bernoulli}(p_e), p_e = 0.2/(1 + \exp(-0.1s)) + 0.2U, U \sim \text{Uniform}[0, 1]$ and $\pi^b(a \mid s) \sim \text{Bernoulli}(p_b), p_b = 0.9/(1 + \exp(-0.1s)) + 0.1U, U \sim \text{Uniform}[0, 1]$. We consider the interacted state-action basis, i.e. fit Q on $s + s * a$ with an intercept. When Q is well-specified, we do nearly exactly recover the right contrast function; although in such a small and well-specified example we do not see benefits of orthogonality.

Table 3: Validation: 1d example, linear FQE vs orthogonal τ estimation. Entries: Mean-squared error of estimated τ function, mean + 1.96 standard errors.

Method (n)	$5.0 \cdot 10^1$	$1.3 \cdot 10^3$	$2.5 \cdot 10^3$	$3.8 \cdot 10^3$	$5.0 \cdot 10^3$
FQE	$4 \cdot 10^{-2} \pm 2 \cdot 10^{-2}$	$2 \cdot 10^{-3} \pm 10^{-3}$	$10^{-3} \pm 10^{-4}$	$10^{-3} \pm 10^{-4}$	$4 \cdot 10^{-4} \pm 10^{-4}$
OrthDiff-Q	$6 \cdot 10^{-2} \pm 3 \cdot 10^{-2}$	$3 \cdot 10^{-3} \pm 10^{-3}$	$2 \cdot 10^{-3} \pm 10^{-3}$	$2 \cdot 10^{-3} \pm 10^{-3}$	$2 \cdot 10^{-3} \pm 10^{-4}$

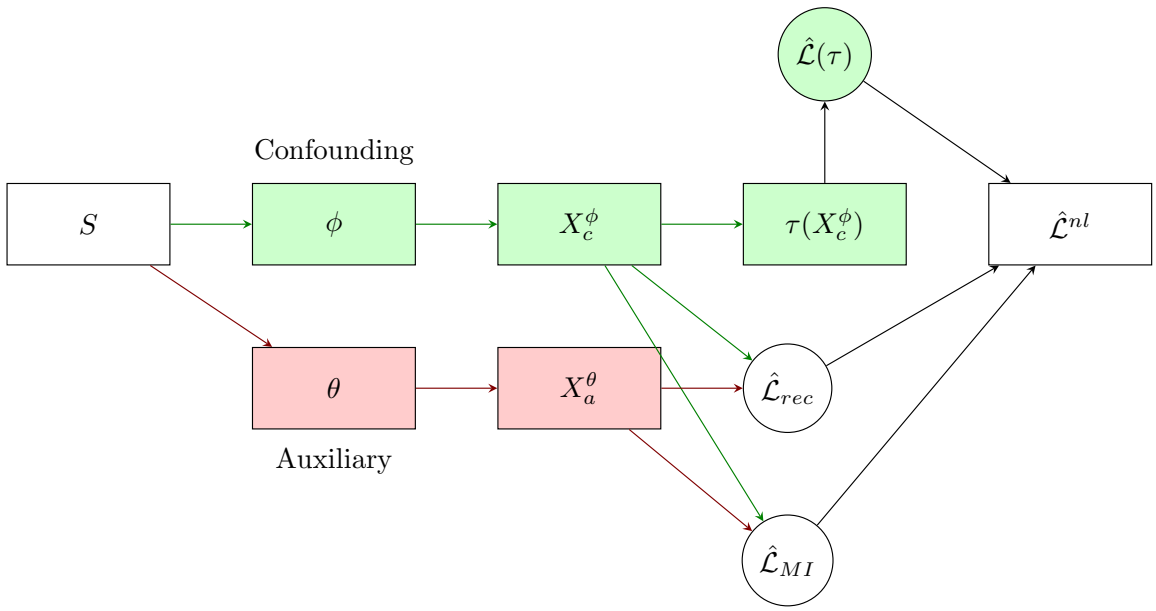


Figure 3: Heuristic neural network architecture diagram for nonlinear mutual information regularization.

Data generating process details on "Adapting to Structure" experiment For the reward-filtered DGP, $|\mathcal{S}| = 100$ though the first 15 dimensions are the reward-relevant sparse component, where ρ is the indicator vector of the sparse support, and $\mathcal{A} = \{0, 1\}$. The reward and states evolve according to $r_t(s, a) = \beta^\top \phi_t(s, a) + a * \sum_{k=1}^{5} s_k/2 + \epsilon_r, \quad s_{t+1}(s, a) = M_a s + \epsilon_s$, satisfying the graphical restrictions of Section 1. Therefore the transition matrices are $M_a = \begin{bmatrix} M_a^{\rho\to\rho} & 0 \\ M_a^{\rho\to\rho_c} & M_a^{\rho_c\to\rho_c} \end{bmatrix}$. We generate the coefficient matrices M_0, M_1 with independent normal random variables $\sim N(0.2, 1)$. The nonzero mean ensures the beta-min condition. We normalize $M_a^{\rho\to\rho}$ to have spectral radius 1, then introduce mild instability in the exogenous component by dividing $M_a^{\rho_c\to\rho_c}$ by 0.8x the largest eigenvalue. Therefore, recovering the sparse component is stable but including distracting dimensions destabilizes. The noise terms are normally distributed with standard deviations $\sigma_s = 0.3, \sigma_r = 0.5$. Features $\phi(s, a) = \langle s, sa, 1\rangle$ are the interacted state-action space. The behavior policy is a mixture of logistic, with coefficients $\sim N(0, 0.3)$, and 20% probability of uniform random sampling. The evaluation policy is logistic, with coefficients $\sim$ Unif$[-0.5, 0.5]$. (We fix the random seed).

Details on nonlinear mutual information extension