Structured Difference-of-Q via Orthogonal Learning

Defu Cao
Department of Computer Science
University of Southern California

Angela Zhou
Department of Data Sciences and Operations
Department of Computer Science
University of Southern California

Abstract

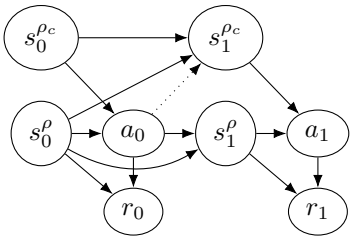
Offline reinforcement learning is important in many settings with available observational data but the inability to deploy new policies online due to safety, cost, and other concerns. Many recent advances in causal inference and machine learning target estimation of "causal contrast" functions such as CATE, which is sufficient for optimizing decisions and can adapt to potentially smoother structure. We develop a dynamic generalization of the R-learner [25, 19] 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

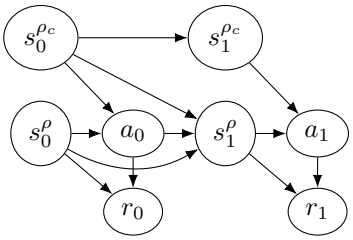
Offline reinforcement learning shares deep connections with causal inference. An extensive literature on causal inference and machine learning establishes methodologies for learning *causal contrasts,* such as the *conditional average treatment effect* (CATE) [35, 9, 17, 15], the covariate-conditional difference in outcomes under treatment and control, which is sufficient for making optimal decisions. A key "inductive bias" motivation is that the causal contrast (i.e. the difference that actions make on outcomes) may be smoother or more structured (e.g., sparser) than the main effects (what happens under either action by itself, Q^π). Methods that specifically estimate these contrast functions could potentially adapt to this favorable structure when it is available. A classically-grounded and rapidly growing line of work on double, orthogonal, or debiased machine learning [16, 4] derives improved estimation procedures for these targets. Estimating the causal contrast can be statistically favorable. 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 $\mathcal{A}$.

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 [38, 37]. For example a certain transition-reward factorization, first studied by [5], admits a sparse Q-function contrast [27]. [42] proposes a variant of the underlying blockwise pattern that also admits sparse optimal Q functions and policies. Figure 1a and Section 1 illustrates how both of these structures have very different conditional independence 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 rather than Q-functions adapts to underlying structure, such as sparsity or smoothness, even when the individual Q-functions are more complex.

Preprint.



(a) Reward-relevant/irrelevant factored dynamics of [42]. The dotted line from a_t to $s_{t+1}^{\rho_c}$ means presence or absence is permitted. Q is sparse in s_0^{ρ} (i.e. doesn't change if $s_0^{\rho_c}$ changes).



(b) Exogenous-Endogenous MDP model of [5]. Q is not sparse in s_0^{ρ} but $Q(s,1) - Q(s,0)$ is.

Table 1: Comparison of Desiderata and Methods.

Features	DROPE [11, 13]	Dyn-R [18]	DAE [27, 28]	FQE [6, 34]	Diff-Q (Ours)
Difference of Q function	✗	✗	✓	≈	✓
Orthogonal estimation	✓	✓	✗	✗	✓
Avoids Multiplied IS	≈	✓	✓	✓	✓
Convex loss	n/a	✓	✗	✓	✓

In this work, we extend the R-learner approach [26, 19] 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 than usual, 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 $O_p(n^{-\frac{1}{2}})$ rates of convergence for estimating the difference of Q functions and attaining the optimal policy value.

Related Work Due to the short workshop format, we have most of the related work in the appendix.

2 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 $t = 1, \ldots, T$ 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 $(S_t, A_t, \ldots)$, lower case letters s, a denote evaluation at a generic value.

The value function is $V_t^{\pi}(s) \coloneqq \mathbb{E}_{\pi}[\sum_{t'=t}^{T} \gamma^{t'-t} R_{t'} \mid S_t = s]$ 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, and satisfies the Bellman evaluation operator:

$$Q_t^{\pi}(s,a) \coloneqq \mathbb{E}_{\pi}[\textstyle\sum_{t'=t}^{T} \gamma^{t'-t} R_{t'} \mid S_t = s, A_t = a] = \mathbb{E}[R_t + \gamma V_{t+1}^{\pi}(S_{t+1}) \mid S_t = s, A_t = a]$$

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 $\|f(X)\|_2 \coloneqq \mathbb{E}[|f(X)|^2]^{1/2}$; expectations and norms are under the observational behavior distribution unless otherwise indicated.

Policy Evaluation (Identification):

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:

$$\tau_t^{\pi}(S_t) \coloneqq Q_t^{\pi}(S_t, 1) - Q_t^{\pi}(S_t, 0). \tag{1}$$

Algorithm 1 Dynamic Residualized Difference-of-Q-Evaluation

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

See the appendix for full details on the derivation. We obtain the identifying moment condition, satisfied by the true difference-of-Q function $\tau_t^\pi(S_t)$:

$$\mathbb{E}[\{R_t + \gamma Q_{t+1}^\pi(S_{t+1}, A_{t+1}) - m_t^\pi(S_t)\} - \{A - \pi_1^b(S_t)\}\tau_t^\pi(S_t) \mid S_t, A_t] = 0 \tag{2}$$

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

$$\begin{aligned} &\tau_t(\cdot) \in \underset{\tau \in \mathcal{G}}{\text{argmin}}\, \mathcal{L}(\tau, \eta), \\ &\mathcal{L}_t(\tau, \eta) \coloneqq \mathbb{E}\left[\left(\{R_t + \gamma Q_{t+1}^\pi(S_{t+1}, A_{t+1}) - m_t^\pi(S_t)\} - \{A - \pi_t^b(1 \mid S_t)\} \cdot \tau(S_t) \right)^2 \right] \end{aligned} \tag{3}$$

Policy optimization. The sequential loss minimization approach also admits an policy optimization procedure. The policy is greedy with respect to the estimated τ_t, which is re-estimated at every timestep. The algorithm statement is deferred to the appendix given the short workshop format. We use a slightly different cross-fitting approach for policy optimization.

3 Analysis

We study the improved statistical rates of convergence from orthogonal estimation for policy evaluation (Theorem 3.1) and show that this implies convergent policy optimization (Theorem 3.2). Theorem 3.1 applies orthogonal statistical learning to our new estimand, for which we establish Neyman-orthogonality. Policy optimization is more challenging; the novelty of Theorem 3.2 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 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 $\mathcal{T}^* f_{t+1}(s, a) = \mathbb{E}[r_t + \max_a f_{t+1}(S_{t+1}, a) \mid s, a]$, $\sup_{f_{t+1} \in \mathcal{F}_{t+1}} \inf_{f_t \in \mathcal{F}_t} \|f_t - \mathcal{T}^* f_{t+1}\|_2^2 \leq \epsilon$.

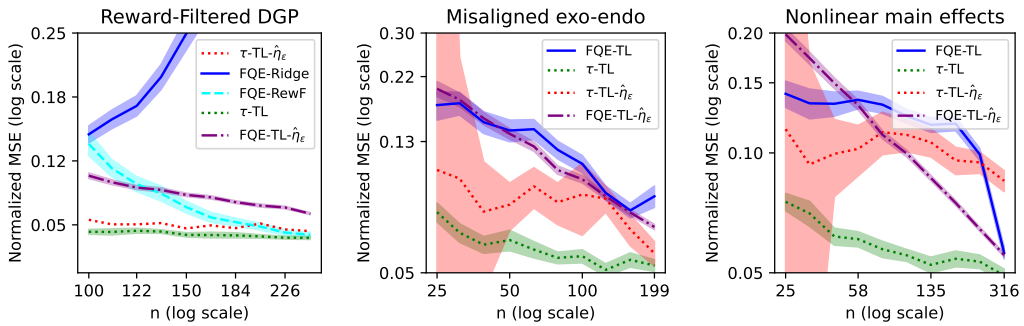


Figure 2: Adapting to structure. Interacted setting where $E[M_1 - M_0] = 0.1 \cdot I$.

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 that is a primitive condition for the product-error rate assumption. Next we establish convergence rates of $\hat{\tau}^\pi$, depending on convergence rates of the nuisance functions. 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 3.1 (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 product RMSE error rates for* π_b, Q *are* $o_p(n^{-\frac{1}{2}})$ *(Assumption 4). Fix the evaluation policy* π^e*. Then, for* $\sigma > 0$, $\|\hat{\tau}_t^{\pi^e} - \tau_t^{0,\pi^e}\|_2 \lesssim n^{-\frac{1}{2}}$.

For example, this 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 $n^{-\frac{1}{2}}$. Methods for estimating Q would require $n^{-\frac{1}{2}}$ convergence.

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 [33]. It is commonly used to relate plug-in estimation error to decision risk.

Assumption 7 (Margin on observational distribution)**.** Let $Q_t^*(s, \pi^*)$ denote the optimal Q function at the optimal action, and a' denote the second-best option, $a' \in \mathcal{A} \setminus \arg\max_a Q_t^*(s, a)$. Assume there exist some 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, $S_t \sim \mu_{\pi_t^b}$.

Next we study policy optimization.

Theorem 3.2 (Policy optimization bound)**.** *Suppose Assumptions 1 to 6 and Assumption 7 (margin assumption holds with* α*). Suppose the product error rate conditions of Assumption 4 hold for each* t *for* $\hat{\underline{\pi}}_{t+1}$*, the data-optimal policies evaluated along the algorithm steps. Then for* $\hat{\pi}_t$*, Theorem 3.1 holds. And,*

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

Experiments Due to space constraints, discussion of experiments is in the Appendix. We also have a nonlinear extension with a deep network architecture and 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. We also add a reconstruction loss function $\hat{\mathcal{L}}_{rec}(\phi, \theta)$ which ensures that these two representations jointly recover the state.

References

[1] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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*.

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

[10] 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*.

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

[12] 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.

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

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

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

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

[17] 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.

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

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

[20] 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.

[21] 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.

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

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

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

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

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

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

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

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

[30] 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.

[31] 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*.

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

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

[34] 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*.

[35] 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.

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

[37] 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*.

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

[39] 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.

[40] 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.

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

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

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:

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

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.

Pan and Schölkopf [27] 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 [28] 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.

A Related work and Additional discussion

There is a large body of work on offline policy evaluation and optimization in offline reinforcement learning [12, 40], including approaches that leverage importance sampling or introduce marginalized versions [11, 32, 13, 20]. 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 [13, 14, 39]. The literature on dynamic treatment regimes (DTRs) studies a method called advantage learning [30], 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.

Advantage functions appear in RL and dynamic treatment regimes [24, 23]. However, policy optimization is hard because *which* contrast it evaluates is time-t policy dependent when optimizing at time t. Our difference-of-Q functions are independent of candidate time-t policies when optimizing at time t.

Pan and Schölkopf [27, 28] estimate the advantage function without estimating Q-functions, but introduces a more difficult nonconvex constraint on the action-average of advantages. The motivations are different; the methods are complementary; we focus on improved statistical guarantees. Farias et al. [7] 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.

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 [21], fitted-Q-iteration [3], or other algorithms.

B 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 3.1 and 3.2 can be obtained via Assumption 5 (well-specification and exact solution so that $\epsilon\left(\tau_t^n, \hat{\eta}\right) = 0$, 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* **??**. *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:*

$$\begin{aligned}\rho_t^{\pi^e}(\hat{\eta}) = B_\tau{}^2 \|(\hat{\pi}_t^b - \pi_t^{b,\circ})^2\|_u + B_\tau \|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{m}_t^{\pi^e} - m_t^{\pi^e,\circ})\|_u \\ + \gamma(B_\tau \|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{Q}_{t+1}^{\pi^e} - Q_{t+1}^{\pi^e,\circ})\|_u + \|(\hat{m}_t^{\pi^e} - m_t^{\pi^e,\circ})(\hat{Q}_{t+1}^{\pi^e} - Q_{t+1}^{\pi^e,\circ})\|_u). \quad (4)\end{aligned}$$

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 2.

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 $\mathrm{star}(\mathcal{F}) = \{cf : f \in \mathcal{F}, c \in [0, 1]\}$. 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. [36]. We can obtain mean-squared error rates for policy evaluation via specializing Theorem 3.1 to the 2-norm and leveraging results from [9].

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 [8]. 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 [29].

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* **??**. *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 1 (Advantage estimation error to policy value via margin.)**.** *Suppose* **??** *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* $b_* > 0$*,*

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

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

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

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* **??**. *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* $\hat{\overline{\pi}}_{t+1}$*, the data-optimal policies evaluated along the algorithm steps, with rates* ρ_t *for each timestep. Suppose that then for* $\hat{\pi}_t$*,* **??** *holds. Denote the slowest such product-error-rate over timesteps as* $\overline{\rho}_{1:T} = \min_t\{\rho_t\}$*. Then,*

$$\|\hat{\tau}_t^{\hat{\overline{\pi}}_{t+1}} - \tau_t^{\circ,\overline{\pi}^*_{t+1}}\|_2 \leq O(n^{-\overline{\rho}_{1:T}}), \text{ and } \left|\mathbb{E}[V_1^{\pi^*}(S_1) - V_1^{\hat{\pi}_{\hat{\tau}}}(S_1)]\right| = O(n^{-\{\overline{\rho}_{1:T}\}\frac{2+2\alpha}{2+\alpha}}). \quad (5)$$

C Proofs

C.1 Preliminaries

Lemma 2 (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.

$$\mathbb{E}[\hat{\mathcal{L}}_t(\tau,\eta)]$$

$$= \mathbb{E}[\Big(\{R_t + \gamma Q^{\pi^e}_{t+1}(S_{t+1},A_{t+1}) - m^\pi_t(S_t)\} \pm \mathbb{E}[\gamma Q^{\pi^e}_{t+1} \mid S_t, \pi^b_t] - \{A - \pi^b_t(1 \mid S_t)\}\cdot\tau(S_t)\Big)^2]$$

$$= \mathbb{E}[\Big(\{R_t + \gamma\mathbb{E}[Q^{\pi^e}_{t+1} \mid S_t,\pi^b_t] - m^\pi_t(S_t)\} - \{A-\pi^b_t(1\mid S_t)\}\cdot\tau(S_t) + \gamma(Q^{\pi^e}_{t+1}(S_{t+1},A_{t+1}) - \mathbb{E}[Q^{\pi^e}_{t+1}\mid S_t,\pi^b_t])\Big)^2]$$

$$= \mathbb{E}[\Big(\{R_t + \gamma\mathcal{T}Q^{\pi^e}_{t+1} - m^\pi_t\} - \{A-\pi^b_t(1\mid S_t)\}\cdot\tau(S_t)\Big)^2]$$

(squared loss of identifying moment)

$$+ \mathbb{E}[\gamma(Q^{\pi^e}_{t+1}(S_{t+1},A_{t+1}) - \mathbb{E}[Q^{\pi^e}_{t+1}\mid S_t,\pi^b_t])^2]$$ (residual variance of $Q_t(s,a) - R_t(s,a)$)

$$+ \mathbb{E}[\Big\{R_t + \gamma\mathbb{E}[Q^{\pi^e}_{t+1}\mid S_t,\pi^b_t] - m^\pi_t(S_t) - \{A-\pi^b_t(1\mid S_t)\}\cdot\tau(S_t)\Big\}\cdot\gamma(Q^{\pi^e}_{t+1}(S_{t+1},A_{t+1}) - \mathbb{E}[Q^{\pi^e}_{t+1}\mid S_t,\pi^b_t])]$$

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 $\nu_t = \hat{\tau}_t - \tau^n_t, \nu^\circ_t = \hat{\tau}_t - \tau^\circ_t$. 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 3 (Universal Orthogonality)**.**

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

Proof of Lemma 3. For brevity, for a generic f, let $\{f\}_\epsilon$ denote $f + \epsilon(f - f^\circ)$. Then the first Frechet derivatives are:

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

$$\frac{d}{d\epsilon_e}\frac{d}{d\epsilon_\tau}\mathcal{L}_t\left(\tilde{\tau},\eta^\circ\right)\left[\eta-\eta^\circ,\tau-\tau\right]\Big|_{\epsilon=0}$$

$$= \mathbb{E}\left[\left(\pi^b_t - \pi^{b,\circ}_t\right)\tau(\tau-\tilde{\tau})(A_t-e_t)\right] + \mathbb{E}\left[\Big\{R + \gamma Q^{\pi^e}_{t+1} - m^{\pi^e,\circ}_t - (A_t-e_t)\Big\}(\tau-\tilde{\tau})\cdot -(e_t - e^\circ_t)\right]$$

$$= 0$$

$$\frac{d}{d\epsilon_{Q_{t+1}}}\frac{d}{d\epsilon_\tau}\mathcal{L}_t\left(\tilde{\tau},\eta^\circ\right)\left[\eta-\eta^\circ,\tau-\tilde{\tau}\right]\Big|_{\epsilon=0}$$

$$= \mathbb{E}[\gamma(Q^{\pi^e}_{t+1} - Q^{\pi^e,\circ}_{t+1})(A_t - \pi^{b,\circ}_t)(\tau_t - \tilde{\tau}_t)]]$$

$$= 0$$

$$\frac{d}{d\epsilon_{m_t}}\frac{d}{d\epsilon_\tau}\mathcal{L}_t\left(\tilde{\tau},\eta^\circ\right)\left[\eta-\eta^\circ,\tau-\tilde{\tau}\right]\Big|_{\epsilon=0}$$

$$= \mathbb{E}[-(m^{\pi^e}_t - m^{\pi^e,\circ}_t)(A_t - \pi^{b,\circ}_t)(\tau_t - \tilde{\tau}_t)]$$

$$= 0$$

□

Lemma 4 (Second order derivatives)**.** *For Q_{t+1}, Q°_{t+1} evaluated at some fixed policy π^e:*

$$D_{\eta_t,\eta_t}\mathcal{L}_t[\hat{\eta}_t - \eta^\circ_t, \hat{\eta}_t - \eta^\circ_t]$$

$$= \mathbb{E}\left[\tau^2_t\left(\hat{\pi}^b_t - \pi^{b,\circ}_t\right)^2\right] + \mathbb{E}\left[(\hat{\pi}^b_t - \pi^{b,\circ}_t)\tau_t(\hat{m}_t - m^\circ_t)\right] + \mathbb{E}\left[(\hat{\pi}^b_t - \pi^{b,\circ}_t)\tau_t\gamma(\hat{Q}_{t+1} - Q^\circ_{t+1})\right]$$

$$- \mathbb{E}\left[(\hat{m}_t - m^\circ_t)\,\gamma\left(\hat{Q}_{t+1} - Q^\circ_{t+1}\right)\right]$$

Proof of Lemma 4. Below, the evaluation policy π^e is fixed and omitted for brevity. Note that

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

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

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

By the chain rule for Frechet differentiation, we have that

$$D_{\eta_t,\eta_t}\mathcal{L}_t[\hat{\eta}_t - \eta_t^\circ, \hat{\eta}_t - \eta_t^\circ] = D_{\pi_t^b,\pi_t^b}\mathcal{L}_t[\hat{\pi}_t^b - \pi_t^{b,\circ}, \hat{\pi}_t^b - \pi_t^{b,\circ}]$$
$$+ D_{m_t,\pi_t^b}\mathcal{L}_t[\hat{\pi}_t^b - \pi_t^{b,\circ}, \hat{m}_t - m_t^\circ] + D_{Q_{t+1},\pi_t^b}\mathcal{L}_t[\hat{\pi}_t^b - \pi_t^{b,\circ}, \hat{Q}_{t+1} - Q_{t+1}^\circ] + D_{m_t,Q_{t+1}}\mathcal{L}_t[\hat{Q}_{t+1} - Q_{t+1}^\circ, \hat{m}_t - m_t^\circ]$$

□

C.3 Proof of sample complexity bounds

Proof of **??**. 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).

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

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

$$\mathbb{E}[V_t^*(s)] - \mathbb{E}[V_t^{\pi_{\hat{\tau}}}(s)] \leq \gamma\mathbb{E}\left[\mathbb{E}_{\hat{\pi}_t}[V_{t+1}^{\pi^*} - V_{t+1}^{\hat{\pi}_{\hat{\tau}}} \mid s]\right] + \mathbb{E}[Q_t^*(s, \pi^*) - Q_t^*(s, \hat{\pi}_{\hat{\tau}})] \tag{6}$$

Assuming bounded rewards and Assumption 9 implies that $P(s_{t+1} \mid s, a) \leq c$, 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:

$$\mathbb{E}_\pi[Q_t^*(s_t, \pi^*) - Q_t^*(s_t, \hat{\pi}_{\hat{\tau}})] \leq c\int \{Q_t^*(s, \pi^*) - Q_t^*(s, \hat{\pi}_{\hat{\tau}}\}\, ds, \tag{7}$$

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. (7).

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

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

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

$$\int \{Q_t^*(s,\pi^*(s)) - Q_t^*(s,\hat{\pi}_{\hat{\tau}})\}\, ds = \int \{Q_t^*(s,\pi^*(s)) - Q_t^*(s,\hat{\pi}_{\hat{\tau}})\}\, \mathbb{I}\left[\mathcal{E}_g\right] ds + \int \{Q_t^*(s,\pi^*(s)) - Q_t^*(s,\hat{\pi}_{\hat{\tau}})\}\, \mathbb{I}\left[\mathcal{E}_g^c\right] ds \tag{8}$$

Assuming boundedness, the bad event occurs with vanishingly small probability $n^{-\kappa}$, which bounds the second term of eqn. (12).

For the first term of eqn. (12), 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 $\left|\tau_t^{\pi_{t+1}^*,\circ}(s)\right| \leq 2Kn^{-b_*}$. And if no mistakes occur, at s the contribution to the integral is 0. Denote the mistake region as

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

Therefore

$$\int \{Q_t^*(s,\pi^*(s)) - Q_t^*(s,\hat{\pi}_{\hat{\tau}})\}\, ds \leq \int_{s\in\mathcal{S}_m} \{Q_t^*(s,\pi^*(s)) - Q_t^*(s,\hat{\pi}_{\hat{\tau}})\}\, \mathbb{I}\left[s \in \mathcal{S}_m\right] \mathbb{I}\left[\mathcal{E}_g\right] ds + O(n^{-\kappa}) \tag{9}$$

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:

$$|Q^*(s,\pi^*) - Q^*(s,\hat{\pi})| \leq |\tau^{\pi_{t+1}^*,\circ}| \leq 2Kn^{-b*}. \tag{10}$$

Therefore,

$$\begin{aligned}
\int \{Q_t^*(s,\pi^*(s)) - Q_t^*(s,\hat{\pi}_{\hat{\tau}})\}\, ds &\leq O(n^{-\kappa}) + Kn^{-b*} \int \mathbb{I}\left[s \in \mathcal{S}_m\right] ds \\
&\leq O(n^{-\kappa}) + (Kn^{-b*})(Kn^{-b*\alpha}) \\
&= O(n^{-\kappa}) + (K^2 n^{-b*(1+\alpha)})
\end{aligned} \tag{11}$$

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

Combining Eqs. (6) and (11), 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:

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

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 [22] for more discussion on the relationship between them.

The results under an integrated risk bound assumption on convergence of τ follow analogously as [31], 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:

$$\left\|Q_t^*\left(S_t,\pi^*\right)-Q_t^*\left(S_t,\hat{\pi}_{\hat{\tau}}\right)\right\|_2 \leq \left\|(Q_t^*\left(S_t,\pi^*\right)-Q_t^*\left(S_t,\hat{\pi}_{\hat{\tau}}\right))\mathbb{I}\left[\mathcal{S}_\epsilon\right]\right\|_2+\left\|(Q_t^*\left(S_t,\pi^*\right)-Q_t^*\left(S_t,\hat{\pi}_{\hat{\tau}}\right))\mathbb{I}\left[\mathcal{S}_\epsilon^c\right]\right\|_2 \tag{12}$$

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

$$\max_a Q_t^*(s,a)-Q^*(s,\hat{\pi}(s)) \leq 2\left|\hat{\tau}^{\hat{\pi}_{t+1}}(s)-\tau^{\pi_{t+1}^*,\circ}(s)\right|. \tag{13}$$

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. (6). □

Proof of Theorem 3.1. 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

$$\begin{aligned}\frac{\lambda}{2}\left\|\nu_t\right\|_2^2 &\leq \mathcal{L}_D(\hat{\tau}_t,\hat{\eta})-\mathcal{L}_D(\tau_t^n,\hat{\eta})-D_{\tau_t}\mathcal{L}_D(\tau_t^n,\hat{\eta})[\nu_t]\\ &\leq \epsilon(\hat{\tau}_t,\hat{\eta})-D_{\tau_t}\mathcal{L}_D(\tau_t^n,\eta^\circ)[\nu_t]\\ &\quad + D_{\tau_t}\mathcal{L}_D(\tau_t^n,\eta^\circ)[\nu_t]-D_{\tau_t}\mathcal{L}_D(\tau_t^n,\hat{\eta})[\nu_t]\end{aligned} \tag{14}$$

We bound each term in turn.

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

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

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 3 and 4 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:

$$\begin{aligned}
\frac{\lambda}{2}\|\nu_t\|_2^2 \leq\ &\epsilon(\hat{\tau}_t, \hat{\eta}) + \|\nu_t\|_{\overline{u}}\|(\tau^\circ - \tau_t^n)\|_u \\
&+ \|\nu_t\|_{\overline{u}}\Bigg({B_\tau}^2\|\left(\hat{\pi}_t^b - \pi_t^{b,\circ}\right)^2\|_u + B_\tau\|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{m}_t - m_t^\circ)\|_u + \gamma B_\tau\|(\hat{\pi}_t^b - \pi_t^{b,\circ})(\hat{Q}_{t+1} - Q_{t+1}^\circ)\|_u \\
&\quad + \gamma\|(\hat{m}_t - m_t^\circ)(\hat{Q}_{t+1} - Q_{t+1}^\circ)\|_u\Bigg)
\end{aligned} \tag{15}$$

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

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

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

$$\frac{\lambda}{2}\|\nu_t\|_2^2 - \frac{\sigma}{4}\|\nu_t\|_{\overline{u}}^2 \leq \epsilon(\hat{\tau}_t, \hat{\eta}) + \frac{1}{\sigma}\left(\|(\tau^\circ - \tau_t^n)\|_u + \rho_t(\hat{\eta})\right)^2 \tag{16}$$

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

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

□

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 6 (Lemma 14 of [9] on local Rademacher complexity decompositions), with high probability 1-ξ, for all $t \in [T]$, and for c_0 a universal constant ≥ 1.

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

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 $\geq 1 - 2\xi$:

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

The ϵ excess risk term in Theorem 3.1 indeed corresponds to one of the loss differences defined here, i.e. $\Delta\mathcal{L}_{D,t}\left(\hat{\tau}_t, \hat{\eta}_S\right) := \epsilon(\hat{\tau}_t^n, \hat{\tau}_t, \hat{h}_S)$. Therefore, applying Theorem 3.1 with $u = \overline{u} = 2$ and $\sigma = \lambda$ with the above bound, and averaging the sample-split estimators, we obtain

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

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

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

Therefore,

$$\frac{\lambda-\epsilon}{4}\|\hat{\tau}_t-\tau_t^\circ\|_2^2 \le c_0 r^2 m^2(1+\frac{1}{\epsilon})\delta_{n/2,\xi}^2+\left(\frac{8+\lambda^2}{4\lambda}+\frac{\epsilon}{4}\right)\|\tau_t^\circ-\tau_t^n\|_2^2+\frac{2}{\lambda}\sum_{s\in\{S,S'\}}\rho_t(\hat{\eta}_s)^2$$

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

$$\begin{aligned}\frac{\lambda}{8}\|\hat{\tau}_t-\tau_t^\circ\|_2^2 &\le c_0 r^2 m^2(1+\frac{8}{\lambda})\delta_{n/2,\xi}^2+\left(\frac{4+\lambda^2}{2\lambda}\right)\|\tau_t^\circ-\tau_t^n\|_2^2+\frac{2}{\lambda}\sum_{s\in\{S,S'\}}\rho_t(\hat{\eta}_s)^2\\ &\le \left(1+\frac{8}{\lambda}+\frac{\lambda}{2}\right)(c_0 r^2 m^2\delta_{n/2,\xi}^2+\|\tau_t^\circ-\tau_t^n\|_2^2+\sum_{s\in\{S,S'\}}\rho_t(\hat{\eta}_s)^2)\end{aligned}$$

and therefore

$$\|\hat{\tau}_t-\tau_t^\circ\|_2^2 \le \left(\frac{8}{\lambda}(1+\frac{8}{\lambda})+4\right)(c_0 r^2 m^2\delta_{n/2,\xi}^2+\|\tau_t^\circ-\tau_t^n\|_2^2+\sum_{s\in\{S,S'\}}\rho_t(\hat{\eta}_s)^2)$$

Taking expectations:

$$\mathbb{E}[\|\hat{\tau}_t-\tau_t^\circ\|_2^2] \le \left(\frac{8}{\lambda}(1+\frac{8}{\lambda})+4\right)(c_0 r^2 m^2\delta_{n/2}^2+\|\tau_t^\circ-\tau_t^n\|_2^2+\max_{s\in\{S,S'\}}\mathbb{E}[\rho_t(\hat{\eta}_s)^2])$$

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 3.2. **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^*.

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

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 **??** 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

$$\|\hat{\tau}_k^{\underline{\hat{\pi}}_{k+1}} - \tau_k^{\circ,\underline{\pi}^*_{k+1}}\| = O(\delta_{n/2} + \|\tau_k^{\circ,\underline{\hat{\pi}}_{k+1}} - \tau_k^{n,\underline{\hat{\pi}}_{k+1}}\|_2) + Kn^{-\mathcal{R}_k}, \tag{17}$$

where

$$\mathcal{R}_k = \min\left(\rho_{k+1}^{(c)} \cdot \frac{2+2\alpha}{2+\alpha},\ \rho_{k+1}^{(\Psi)} \cdot \frac{2+2\alpha}{2+\alpha},\ -\{\min_{k' \geq k+1}(\rho_{k'}^{(c)}, \rho_{k'}^{(\Psi)})\} \cdot \frac{2+2\alpha}{2+\alpha}^{T-k'}\right). \tag{18}$$

And therefore, applying **??**, that

$$\left|\mathbb{E}[V_k^{\pi^*} - V_k^{\hat{\pi}_{\hat{\tau}}}]\right| = O(n^{-\min\{\rho_k^{(c)}, \rho_k^{(\Psi)}\}\frac{2+2\alpha}{2+\alpha}}) + o(n^{-\min\{\rho_k^{(c)}, \rho_k^{(\Psi)}\}\frac{2+2\alpha}{2+\alpha}}). \tag{19}$$

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:

$$\|\hat{\tau}_t^{\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\pi}^*_{t+1}}\| \leq \|\hat{\tau}_t^{\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\hat{\pi}}_{t+1}}\| + \|\tau_t^{\circ,\underline{\hat{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\pi}^*_{t+1}}\| \tag{20}$$

The first term is the policy evaluation estimation error, and under the product error rate assumptions , Theorem 3.1 and **??** 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 **??** 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. (6), we have that:

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

Decompose:

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

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

$$\begin{aligned}
&\|Q_t^{\underline{\pi}^*_{t+1}}(s,a) - Q_t^{\hat{\underline{\pi}}_{t+1}}(s,a)\| \\
&= \|\mathbb{E}_{\pi_t^a}[V_{t+1}^{\underline{\pi}^*_{t+1}} - V_{t+1}^{\hat{\underline{\pi}}_{t+1}} \mid S_t]\| \\
&\leq \|\mathbb{E}_{\pi_t^a}[V_{t+1}^{\underline{\pi}^*_{t+1}} - V_{t+1}^{\hat{\pi}_{t+1},\underline{\pi}^*_{t+2}} \mid S_t]\| + \|\mathbb{E}_{\pi_t^a}[V_{t+1}^{\hat{\pi}_{t+1},\underline{\pi}^*_{t+2}} - V_{t+1}^{\hat{\underline{\pi}}_{t+1}} \mid S_t]\| \\
&= \|\mathbb{E}_{\pi_t^a}[Q_{t+1}^{\underline{\pi}^*_{t+2}}(S_{t+1}, \pi^*_{t+1}) - Q_{t+1}^{\underline{\pi}^*_{t+2}}(S_{t+1}, \hat{\pi}_{t+1}) \mid S_t]\| + \gamma\|\mathbb{E}_{\pi_t^a}[\mathbb{E}_{\hat{\underline{\pi}}_{t+1}}[V_{t+2}^{\underline{\pi}^*_{t+2}} - V_{t+2}^{\hat{\underline{\pi}}_{t+2}} \mid S_t]]\| \qquad (21) \\
&\leq C_\infty \left\{ \int (Q_{t+1}^{\underline{\pi}^*_{t+2}}(s, \pi^*_{t+1}) - Q_{t+1}^{\underline{\pi}^*_{t+2}}(s, \hat{\pi}_{t+1}))^2 ds \right\}^{1/2} + \gamma\|\mathbb{E}_{\pi_t^a}[\mathbb{E}_{\hat{\underline{\pi}}_{t+1}}[V_{t+2}^{\underline{\pi}^*_{t+2}} - V_{t+2}^{\hat{\underline{\pi}}_{t+2}} \mid S_t]]\| \qquad (22)
\end{aligned}$$

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 **??** 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

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

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

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

Combining with the previous analysis, we obtain:

$$\begin{aligned}
\|\hat{\tau}_t^{\hat{\underline{\pi}}_{t+1}} - \tau_t^{\circ,\underline{\pi}^*_{t+1}}\|_2^2 \leq O(\delta_{t,n/2}^2 + \|\tau_t^{\circ,\hat{\underline{\pi}}_{t+1}} - \tau_t^{n,\hat{\underline{\pi}}_{t+1}}\|_2^2) + O(n^{-\min\left\{\rho_{t+2}^{(c)}, \rho_{t+2}^{(\Psi)}, \mathcal{R}_{t+2}\right\}\frac{2+2\alpha}{2+\alpha}})\} \\
+ \frac{(1-\gamma^{T-t-1})}{1-\gamma} C_\pi (T-t-1)\{O(n^{-\kappa}) + Kn^{-\min\{\rho_{t+1}^{(c)}, \rho_{t+1}^{(\Psi)}, \mathcal{R}_{t+1}\}\frac{2+2\alpha}{2+\alpha}}\} \qquad (23)
\end{aligned}$$

from eqn. (21) 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 5 (Lemma 18 of [19]). *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 6 (Lemma 14 of [9], see also results on local Rademacher complexity [36]). *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 $\delta_n^2 \geq \frac{4d\log(41\log(2c_2n))}{c_2 n}$ be any solution to the inequalities:*

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

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 $1 - c_5\exp\left(c_6 n\delta_n^2\right)$,

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

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.

E.1 Omitted details

1D validation. In a very small 1D toy example (Sec 5.1, [13]) we validate our method. See Appendix E of the appendix for more details.

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 [37, 38, 5]. 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 d.g.p.

We investigate the benefits of targeting estimation of the difference-of-Qs in two different graphical substructures, replicated in Section 1 and Figure 1a, proposed in Zhou [42], Dietterich et al. [5]. 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 [42]. In the 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 $r_t(s,a) = \beta^\top\phi_t(s,a) + a * \sum_{k=1}^5 s_k/2 + \epsilon_r$, $s_{t+1}(s,a) = M_a s + \epsilon_s$, satisfying the graphical restrictions of Figure 1a. 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).

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 [42]. Next we have two variants of our framework: in dotted green τ-TL which uses reward-based thresholding to estimate τ

Table 2: Performance comparison on different sample numbers under the nonlinear setting.

Method (n)	100	200	400	600	800
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	2.104 ± 2.392	0.280 ± 0.222	1.179 ± 1.840	1.286 ± 1.123	2.342 ± 1.812

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 solve the loss function minimization exactly 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 (we divide by the square of the range of the true difference of Qs), and the x axis is the number of episodes, on a log scale. First on the left, we consider the previously mentioned reward-filtered DGP. The tailored method of [41] is well-specified. For the reward-filtered DGP, we compare against 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 the case of nuisance function estimation error, as indicated by the red-dotted line where we plug-in quarter-root consistent estimates. (The additional sample splitting leads to transient small-data issues but does not affect the rate of convergence.) 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 [5] (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 regularizer based on mutual information regularization. We use a cartpole-with-distractors environment from Hao et al. [10], which appends additional autoregressive noise to the state in CartPole [2]. Hao et al. [10] focuses on off-policy evaluation with abstractions, so the methods are not comparable.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 how the identification argument extends to the stationary discounted infinite-horizon setting.

We explore the use of mutual information (MI) as a regularization term to optimize our loss function. 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

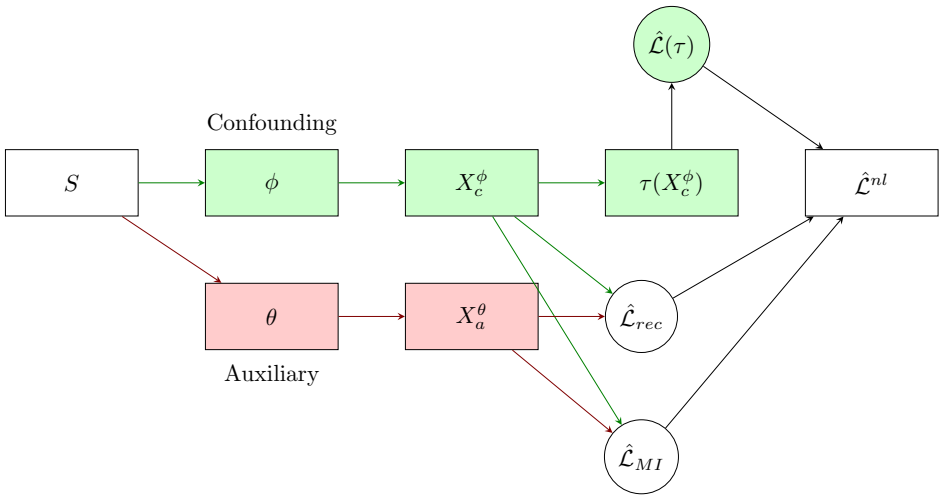


Figure 3: Heuristic neural network architecture diagram for nonlinear mutual information regularization.

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 [1]. (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 functional form access to probability densities, though only samples from the joint distribution in ML-based methods are available. MINE uses these samples.

We illustrate how our method can improve upon naive FQE (learned with neural nets) learned on the full state space. We compare 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-dim state space without distractors. We compare to our DiffQ estimation with neural nets, and a regularized version. Model selection in offline RL is somewhat of an open problem, we leave this for future work.

1d validation example (??) Following the specification of [13, 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.

Details on nonlinear mutual information extension