Reward-Relevance-Filtered Linear Offline Reinforcement Learning

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

Abstract

This paper studies offline reinforcement learning with linear function approximation in a setting with decision-theoretic, but not estimation sparsity. The structural restrictions of the data-generating process presume that the transitions factor into a sparse component that affects the reward and could affect additional exogenous dynamics that do not affect the reward. Although the minimally sufficient adjustment set for estimation of full-state transition properties depends on the whole state, the optimal policy and therefore state-action value function depends only on the sparse component: we call this *causal/decision-theoretic sparsity*. We develop a method for *reward-filtering* the estimation of the state-action value function to the sparse component by a modification of thresholded lasso in least-squares policy evaluation. We provide theoretical guarantees for our reward-filtered linear fitted-Q-iteration, with sample complexity depending only on the size of the sparse component.

1 Introduction

Offline reinforcement learning, learning to make decisions from historical data, is necessary in important application areas such as healthcare, e-commerce, and other real-world domains, where randomized exploration is costly or unavailable. It requires certain assumptions such as full observability and no unobserved confounders. This motivates, especially in the era of big data, collecting as much information as possible about the environment into the state variable. On the other hand, common sensing modalities by default capture not only information that can be affected by an agent's actions, but also information about the environment that is unaffected by an agent's actions. For example, in robotics applications, the dynamics of clouds moving in the sky is a separate process that does not affect, nor is affected by, agents' actions, and does not affect agent reward. Given the overall high variance of learning offline, removing such exogenous information can help improve policy information and optimization, while recovering a minimally sufficient state variable for the optimal policy can reduce vulnerability to distribution shifts.

Though various combinations of relevance/irrelevance are possible for rewards and actions, as has been recognized in a recent work, most works methodologically impose statistically difficult conditional independence restrictions with variational autoencoders that lack strong theoretical computational/statistical guarantees. Other approaches suggest simpler variable screening, but without discussion of underlying signal strength assumptions, or tradeoffs in downstream estimation and value under potential false negatives/positives, and without guarantees. To bridge between these methods, we focus on a model with linear function approximation, a popular structural assumption in the theoretical literature, and develop methods based on thresholded LASSO regression, connecting classical statistical results to new decision-theoretic notions of sparsity introduced by these causal decompositions of reward/action ir/relevance. In particular, we focus on a particular decomposition: the transitions factor into a sparse component that affects the reward, with dynamics that can affect the next timestep's sparse component and an exogenous component. The exogenous component does

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume TBD. Copyright 2024 by the author(s).

not affect the reward or sparse component. A toy example of such a setting is controlling a boat with an image representation of the state environment: actions affect navigation locally and also propagate ripples leaving the boat. Though these ripples evolve under their own dynamics, they themselves do not affect local control of the boat or rewards. Our structural assumptions, though restrictive, still surface what we call "decision-theoretic, but not estimation sparsity": that is, the minimally sufficient causal adjustment set to predict transition probabilities requires the full state variable, but the optimal policy only depends on the sparse component.

The contributions of our work are as follows: under our structural assumptions, we develop methodology for filtering out exogenous states based on support recovery via thresholded lasso regression for the rewards, and linear estimation on the recovered support for the q function via least-squares policy evaluation/fitted-Q-iteration (FQI). We prove predictive error guarantees on the q function estimation, and correspondingly on the optimal policy, showing how the optimal policy now depends on the dimensionality of the sparse component, rather than the full ambient dimension.

2 Preliminaries

We consider a finite-horizon Markov Decision Process on the full-information state space comprised of a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, P, \gamma, T)$ of states, actions, reward function $r(s,a)$, transition probability matrix P, $\gamma < 1$ discount factor, and time horizon of T steps, where $t = 1, \ldots, T$. We let the state spaces $\mathcal{S} \subseteq \mathbb{R}^d$ be continuous, and assume the action space $\mathcal{A}$ is finite: $\phi(s,a)$ denotes a (known) feature map. A policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$ maps from the state space to a distribution over actions, where $\Delta(\cdot)$ is the set of distributions over $(\cdot)$, and $\pi(a \mid s)$ is the probability of taking action a in state s. Since the optimal policy in the Markov decision process is deterministic, we also use $\pi(s) \in \mathcal{A}$ for deterministic policies, to denote the action taken in state s. The policy and MDP $\mathcal{M}$ induce a joint distribution P_π where $P_\pi(a_t \mid s_{0:t}, a_{0:t-1}) = \pi(a_t \mid s_t)$ and $P_\pi(s_{t+1} \mid s_{0:t}, a_{0:t}) = P(s_{t+1} \mid a_t, s_t)$, the transition probability.

The value function is $v_t^\pi(s) = \mathbb{E}_\pi[\sum_{t'=t}^T \gamma^{t'-t} r_{t'} \mid s]$, where $\mathbb{E}_\pi$ denotes expectation under the joint distribution induced by the MDP $\mathcal{M}$ running policy π. The state-action value function, or q function is $q_t^\pi(s) = \mathbb{E}_\pi[\sum_{t'=t}^T \gamma r_{t'} \mid s, a]$. These satisfy the Bellman operator, e.g. $q_t^\pi(s,a) = r(s,a) + \gamma \mathbb{E}[v_{t+1}^\pi(s_{t+1}) \mid s, a]$. The optimal value and q-functions are v^*, q^* correspond to the optimal policy and optimal action, respectively. We focus on the offline reinforcement learning setting where we have access to a dataset of n offline trajectories, $\mathcal{D} = \{(s_t^{(i)}, a_t^{(i)}, s_{t+1}^{(i)})_{t=1}^T\}_{i=1}^n$, where actions were taken according to some behavior policy π_b. We assume throughout that the underlying policy was stationary, i.e. offline trajectories (drawn potentially from a series of episodes) that are independent.

Linearity Throughout this paper, we focus on linear Markov decision processes. Let the feature mapping be denoted $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$. We assume the reward function and value functions are linear in ϕ.

Assumption 1 (Linear MDP)**.** *Assume that both the rewards and transitions are linear functions (possibly with different parameters):*

$$\begin{aligned} r_t(s,a) &= \beta_t \cdot \phi(s,a), \\ q_t^\pi(s,a) &= \theta_t^\pi \cdot \phi(s,a), \\ P_t(\cdot \mid s,a) &= \mu_t \phi(s,a), \forall t \end{aligned}$$

The theoretical analysis of reinforcement learning typically assumes that the reward function is known, since noise in rewards leads to lower-order terms in the analysis. However, in our setting, we will leverage *sparsity of the rewards* to consider minimal state space representations (and adaptive model selection) which affect first-order terms in the analysis.

Linear Bellman completeness is the assumption that for any linear function $f(s,a) := \theta^\top \phi(s,a)$, the Bellman operator applied to $f(s,a)$ also returns a linear function with respect to ϕ. (It is an equivalent assumption but generalizes more directly to potential nonlinear settings).

Definition 1 (Linear Bellman Completeness)**.** *the features ϕ satisfy the linear Bellman completeness property if for all $\theta \in \mathbb{R}^d$ and $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [T]$, there exists $w \in \mathbb{R}^d$ such that:*

$$w^\top \phi(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} \theta^\top \phi(s', a').$$

As w depends on θ, we use the notation $\mathcal{T}_h : \mathbb{R}^d \mapsto \mathbb{R}^d$ to represent such a w, i.e., $w := \mathcal{T}_h(\theta)$ in the above equation. Note that the above implies that $r(s,a)$ is in the span of ϕ (to see this, take $\theta = 0$). Furthermore, it also implies that $q_h^\star(s,a)$ is linear in ϕ, i.e., there exists $\theta_h^\star$ such that $q_h^\star(s,a) = (\theta_h^\star)^\top \phi(s,a)$.

We let $\rho \in [d]$ denote an index set. We use the superscript $(\cdot)^\rho$ to denote subindexing a (random) vector by the index set (since time is the typical subscript), i.e. s^ρ is the subvector of state variable according to dimensions $\rho, s^\rho = \{s_k\}_{k\in\rho}$. We also introduce a new notion of *extension* of a subvector s^ρ to the ambient dimension, i.e. $\breve{s}^{[\rho]} = s_k$ if $k \in \rho$ and 0 otherwise, which makes it easier, for example, to state equivalence of generic q functions comparing full-dimensional states vs. the extension of sparse subvectors to the full-dimensional space, denoted $\breve{q}$.

3 Related work

Our work is related to sparse offline reinforcement learning, LASSO regression for variable selection, and approaches for leveraging causal structure in reinforcement learning to remove important information. We describe each of these in turn.

Structure in offline reinforcement learning. [Hao et al., 2021] studies LASSO estimation for fitted-q-evaluation and interation, and also suggests thresholded LASSO. Although we also use thresholded LASSO, our method is quite different because we directly impose the sparsity structure induced by reward-relevance into estimation of the q function, because the optimal policy is sparse. An emerging line of work identifies causal decomposition of state variables into reward-relevant/reward-irrelevant/controllable components (or variations thereof) [Dietterich et al., 2018, Wang et al., 2022b, Wang et al., Zhang et al., 2020, Seitzer et al., 2021, Efroni et al., 2021]. Methodologically, these works regularize representation learning such as with variational autoencoders towards conditional independence (which generally lacks theoretical guarantees) [Dietterich et al., 2018, Wang et al., 2022a, Seitzer et al., 2021], or assume specific structure such as block MDPs with deterministic latent dynamics emitting high-dimensional observations [Efroni et al., 2021], or require auxiliary non-standard estimation [Lamb et al., 2022]. Our model somewhat resembles the exogenous-endogenous decomposition of [Dietterich et al., 2018], but swaps cross-dependence of exogenous and endogenous components: this gives *different* conditional independence restrictions directly admits sparse learning. Overall, the main simplification of our model relative to these is that rewards do not depend on the exogenous component. The most methodologically related work is that of [Efroni et al., 2022], which studies sparse partial controllability in the linear quadratic regulator; although they also use thresholded LASSO, they consider online control under a different quadratic cost, focus on controllability (action-relevance), and consider entrywise regression of matrix entries.

Variable selection via LASSO. There is an enormous literature on LASSO. We quickly highlight only a few works on thresholded LASSO. [Meinshausen and Yu, 2009] studies model selection properties of thresholded LASSO under a so-called "beta-min" condition, i.e. an assumed lower bound on the smallest non-zero coefficient and gives an asymptotic consistency result. [Zhou, 2010] also studies thresholded LASSO, while [Van de Geer et al., 2011] studies adaptive and thresholded LASSO. For simplicity, we focus on high-probability guarantees under the stronger beta-min condition. But stronger guarantees on thresholded LASSO can easily be invoked instead of the ones we use here. See [Bühlmann and Van De Geer, 2011] as well.

In a different context, that of single-timestep causal inference, [Shortreed and Ertefaie, 2017] proposes the "outcome-adaptive" lasso which adds a coefficient penalty to estimation of the propensity score based on the inverse-strength of coefficients of the outcome model, to screen out covariates unrelated to both exposure and outcome. We are broadly inspired by the idea to encourage sparsity in one model (in our setting, the q-function) based on sparse estimation of another (the reward function). Note, however, that the outcome-adaptive lasso is not applicable to enforce this specific structure.

Our work. Even under our simpler model, leveraging classical results from the sparse regression literature sheds light on different approaches that have already been proposed. For example, Wang et al. [2022b] proposes a variable screening method based on independence testing, which performs better for variable selection than a previous regularization-based method [Wang et al.]. The improvement of thresholding procedures upon regularized LASSO for support recovery is classically well known [Bühlmann and Van De Geer, 2011]. The tighter analysis of thresholded lasso also sheds light on implicit signal strength assumptions and tradeoffs of false positives for downstream policy value.

Overall, relative to works on exogenous structure in reinforcement learning via representation learning, we connect to a classical literature on sparse regression with provable guarantees. On the other hand, relative to an extensive literature on LASSO, the reinforcement learning setting imposes different decision-theoretic desiderata, such that the optimal policy is sparse (hence q-function) even when from

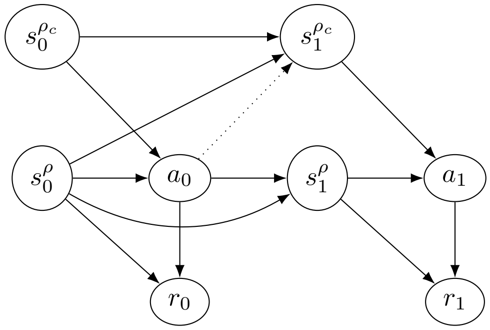


Figure 1: Reward-relevant/irrelevant factored dynamics. The dotted line from a_t to $s_{t+1}^{\rho_c}$ indicates the presence or absence is permitted in the model.

a pure estimation perspective, estimating the transitions are not.

4 Structure

We describe the conditional independence and other restrictions that characterize our filtered reward-relevant model. Let $\rho \subseteq [d]$ denote the supported set of *reward-relevant and endogenous* states. Let $|\rho|$ be the size of the support.

Assumption 2 (Blockwise independent design)**.** $s_t^{\rho} \perp\!\!\!\perp s_t^{\rho_c} \mid s_{t-1}, a_{t-1}$

Assumption 3 (Reward-irrelevant decomposition)**.** *Assume that $R(s,a) = R(\tilde{s},a)$ when $s^{\rho} = \tilde{s}^{\rho}$, and that next-time-step endogenous states are independent of prior exogenous states given prior endogenous states and action:*

$$s_{t+1}^{\rho} \perp\!\!\!\perp s_t^{\rho_c} \mid s_t^{\rho}, a_t \tag{1}$$

The conditional independence restriction implies that $P(s_{t+1}^{\rho} \mid s_t, a_t) = P(s_{t+1}^{\rho} \mid s_t^{\rho}, a_t)$.

Even under these restrictions on the data structure, we can surface a nontrivial qualitative distinction between estimation and decision-making, driven by this causal structure, which we call "causal sparsity" for short. Although the minimal sufficient adjustment set for estimating the entire-state transitions is the non-sparse union of s^{ρ}, s^{ρ_p}, our next results establish that the optimal decision policy is sparse, and hence our thresholded lasso method depends on the sparse component alone.

Note that this decomposition differs from the exogenous-endogenous decomposition in [Dietterich et al., 2018] because our sparse component can affect the exogenous component; but not the other way around – in our model, the exogenous component does not affect the endogenous component.

Let β be the parameter for the q function, and θ be the parameter for the reward function. We let $\sigma_r, \sigma_\theta, \sigma_{r+\gamma q}$ denote the subgaussian parameters of the reward-variance, the Bellman-target, and the transitions, respectively.

Interpreting Assumption 3. For example, consider linear dynamics (with exogenous noise) in an interacted model, i.e. $s_{t+1}(s,a) = M_a s + \epsilon$ for $M_a \in \mathbb{R}^{d \times d}$. Then M_a is a block matrix and it satisfies Assumption 3 if, assuming without loss of generality, that the coordinates are ordered such that the first ρ reward-supported components are first,

$$\begin{aligned} & s_{t+1}(s,a) = M_a s + \epsilon, \\ & \text{where } M_a = \begin{bmatrix} M_a^{\rho \to \rho} & 0 \\ M_a^{\rho \to \rho_c} & M_a^{\rho_c \to \rho_c} \end{bmatrix}. \end{aligned}$$

In particular, the block matrix $M_a^{\rho_c \to \rho} = 0$.

We can also specify a corresponding probabilistic model. Let $P_a(s_{t+1} \mid s_t)$ denote the a-conditioned transition probability, and suppose $P_a(s_{t+1} \mid s_t) \sim N(\mu_a, \Sigma_a)$, and that $P_a(s_{t+1} \mid s_t)$ is partitioned (without loss of generality) as $P_a(s_{t+1}^{\rho}, s_{t+1}^{\rho_c} \mid s_t^{\rho}, s_t^{\rho_c})$. Then by Assumption 3

$$P_a(s_{t+1}^{\rho} \mid s_t^{\rho}) \overset{D}{=} P_a(s_{t+1}^{\rho} \mid s_t^{\rho}, s_t^{\rho_c}) \sim N(\mu_a^{\rho}, \Sigma_a^{\rho,\rho}).$$

where the first equality in distribution follows from the conditional independence restriction of Assumption 3 and the parameters of the normal distribution follow since marginal distributions of a jointly normal random variable follow by subsetting the mean vector/covariance matrix appropriately.

Remark 1. *Similar to previous works studying similar structures, we assume this structure holds. If it may not, we could use model selection methods [Lee et al., 2022]: if we incorrectly assume this structure, we would obtain a completeness violation; so the model selection method's oracle inequalities would apply and be rate-optimal relative to non-sparse approaches. We emphasize that we don't posit this method as a general alternative to general sparsity, but rather as a simple principled approach to estimate in settings with this exogenous structure.*

4.1 Implications for decisions

We characterize important structural properties under the endogenous-exogenous assumption. Under Assumptions 1 and 3, the optimal policy is sparse.

Proposition 1 (Sparse optimal policies)**.** *When $s_t^{\rho} = \tilde{s}_t^{\rho}$, $\pi_t^*(s_t) = \tilde{\pi}_t^*(\tilde{s}_t)$.*

Proposition 1 is the main characterization that motivates our method. Even though the estimation of *transitions* are not sparse, the *optimal* q- and value functions are sparse.

Although well-specification/realizability does not imply Bellman completeness of a function class in general, the reward-sparse linear function class is Bellman-complete for q functions as well. Let $\mathcal{F}_t^\rho$ denote the true sparse function classes $\mathcal{F}_t^\rho = \{\beta \in \mathbb{R}^d \colon \beta_j = 0, j \in \rho\}$.

Proposition 2 (Reward-sparse function classes are Bellman-complete.)**.** *Let $r^\rho(s,a)$ be the ρ-sparse reward function. Let $\breve{q} \in \breve{\mathcal{Q}}$ be the extension of ρ-sparse q functions to the full space, i.e. where $\breve{\mathcal{Q}}$ is the space of functions that are zero outside the support ρ.*

Then: $\sup_{\breve{q}_{t+1} \in \breve{\mathcal{Q}}_{t+1}} \inf_{q_t \in \breve{\mathcal{Q}}_t} \|q_t - \mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2 = 0$

5 Method

Based on the posited endogenous-exogenous structure, the sparsity in the linear rewards is the same sparsity pattern as the optimal value function. Notably, the transitions are not sparse unless only regressing on the endogenous states alone. In our method, we first run thresholded LASSO on rewards to recover the sparse support. Then we fit the q function via ordinary least squares as the regression oracle in least-squares policy evaluation/iteration on the estimated support. We describe each of these components in turn; thresholded LASSO, and fitted-Q-evaluation, before describing our specific method in more detail.

Our main estimation oracle of interest is a variant of *thresholded LASSO*, described in Algorithm 1. We are not limited to thresholded lasso – we could develop analogous adaptations of any method that performs well for support recovery. We simply require finite-sample prediction error guarantees, high probability inclusion of the entire support, and bounds on the number of false positives.

Algorithm 1 Thresholded LASSO

```
Input: (standardized mean-zero and unit variance) covariate matrix $X$, outcome vector $Y$, from data-generating process where $y = w^\top x + \epsilon$.
Obtain an initial estimator $w_{\text{init}}$ using the Lasso.
Let $\hat{\rho} = \{j : w^j_{\text{init}} > \tau_0\}$.
Compute ordinary least squares restricted to $\hat{\rho}$:
    $$\hat{w}^\rho = (X^T_{\hat{\rho}_k} X_{\hat{\rho}})^{-1} X^T_{\hat{\rho}} Y.$$
```

Algorithm 2 Reward-Filtered Fitted Q Iteration

```
At timestep $t = T$ :
Run thresholded LASSO (Algorithm 1) on $r_T$ and obtain sparse support $\hat{\rho}_T$.
$\pi^*_T(s^{\hat{\rho}_T}) = \arg\max_a q_T(s^{\hat{\rho}_T}, a)$.
for timestep $t = T-1, \ldots, 1$ do
    Run thresholded LASSO (Algorithm 1) on $r_t$. Obtain sparse support $\hat{\rho}_t$.
    Compute Bellman target
        $$y_t = r_t + \gamma \mathbb{E}_{\pi^{*;\rho}_{t+1}}[q_{t+1}(s_{t+1}, a_{t+1})].$$
    Fit Bellman residual restricted to $\hat{\rho}_t$.
        $$\widetilde{\beta}_t \in \arg\min_{\beta \in \mathbb{R}^p} \{\tfrac{1}{2}\mathbb{E}_n[(\beta^\top \phi_t - y_t)^2] \colon \beta_j = 0,\ j \in \hat{\rho}_t^c\}$$
    $\pi^*_t(s^\rho) = \arg\max_a q_t(s^\rho, a)$.
end for
```

Fitted-Q-Iteration Linear fitted-q-evaluation, equivalent to offline least-squares policy evaluation, [Ernst et al., 2006, Le et al., 2019, Nedić and Bertsekas, 2003, Duan et al., 2020], and fitted-Q-iteration [Chen and Jiang, 2019, Duan et al., 2021] successively approximate $\hat{q}_t$ at each time step by minimizing an empirical estimate of the Bellman error:

$$\begin{aligned} y_t(q) &:= r_t + \max_{a'} \left[q(s_{t+1}, a')\right], \\ q_t(s,a) &= \mathbb{E}[y_t(q_{t+1}) | s_t = s, a_t = a], \\ \hat{q}_t &\in \arg\min_{q_t \in \mathcal{Q}} \mathbb{E}_{n,t}[(y_t(\hat{q}_{t+1}) - q_t(s_t, a_t))^2]. \end{aligned}$$

The Bayes-optimal predictor of y_t is the true q_t function, even though y_t is a stochastic approximation of q_t that replaces the expectation over the next-state transition with a stochastic sample thereof (realized from data).

Our method Our algorithm, described in Algorithm 2, is a natural modification of these two ideas. At the last timestep, we simply run thresholded lasso on the rewards and set the optimal policy to be greedy with respect to the sparsely-supported reward. At earlier timesteps, we first run thresholded lasso on the rewards and recover an estimate of the sparse support, ρ_t. Then, we fit the Bellman residual $(r_t + \mathbb{E}_{\pi^{*,\rho}_{t+1}}[q_{t+1}(s_{t+1}, a_{t+1})] - q_t(s_t, a_t))^2$ over linear functions of ϕ_t that are supported on ρ_t. That is, we use the sparse support estimated from rewards only in order to sparsely fit the q_t function. Again we set the optimal policy to be greedy with respect to the sparse q_t function.

Why not simply run thresholded LASSO fitted-Q-iteration? Lastly, we provide some important motivation by outlining potential failure modes of simply applying thresholded lasso fitted-Q-iteration (without specializing to the endogenous-exogenous structure here). The first iteration (last timestep), $q_T = R_T$. So thresholded regression at last timestep is analogous to thresholded reward regression. Note that if reward regression succeeds at time T, then we are integrating a dense measure against the sparse function V_T. On the other hand, mistakes in time T will get amplified (i.e. upboosted as "signal" by the dense transition measure). Our reward-thresholded LASSO will not accumulate this error based on the structural assumptions. Without these structural assumptions, it would be unclear whether the rewards are truly dense or whether the dense transitions are amplifying errors in support recovery on the rewards.

6 Analysis

We show a predictive error bound, approximate Bellman completeness under the strong-signal support inclusion of thresholded LASSO, and improvement in policy value. The main technical contribution of our work is the finite-sample prediction error bound for the reward-thresholded fitted-Q-regression. Typical prediction error analyses of thresholded lasso do not directly apply to our setting, where we recover the support from the reward and apply it directly to the q-function estimation. The key observation is that the two regressions share covariance structure and some outcome structure in part. Given this result on the finite-sample prediction error and high-probability inclusion of high-signal sparse covariates, since fitted-Q-evaluation analysis uses prediction bounds on regression in a black-box way, we immediately obtain results on policy value. See [Bühlmann and Van De Geer, 2011, Ariu et al., 2022, Zhou, 2010] for discussion of analysis of thresholded LASSO.

6.1 Preliminaries: standard convergence results for thresholded LASSO

Let $x_t = \phi(s_t, a_t)$ denote regression covariates, with y_t the Bellman residual; in this statement we drop the timestep for brevity and let (X, Y) denote the data matrix and outcome vector, e.g. at a given timestep concatenated over trajectories. Our first assumption is that transition probabilities are time-homogeneous.

Assumption 4. *Time-homogeneous transitions.*

Next we define problem-dependent constants used in the analysis, assumptions, and statements.

Definition 2 (Problem-dependent constants.)**.** *For $a \geq 0$, define*

$$\lambda_{\sigma,a,d} := \sigma\sqrt{1+a}\sqrt{2\log p/n}, \tag{2}$$

$$\mathcal{E}_a := \left\{\epsilon : \left\|X^T\epsilon/n\right\|_\infty \leq \lambda_{\sigma,a,p}\right\}. \tag{3}$$

$\lambda_{\sigma,a,d}$ bounds the maximum correlation between the noise and covariates of X and $\mathcal{E}_a$ is a high probability event where $P(\mathcal{E}_a) \geq 1 - \left(\sqrt{\pi\log p}p^a\right)^{-1}$ when X has column ℓ_2 norms bounded by $\sqrt{n}$. Let $\rho_0 \leq s$ be the smallest integer such that:

$$\textstyle\sum_{i=1}^p \min\left(\beta_i^2, \lambda^2\sigma^2\right) \leq \rho_0\lambda^2\sigma^2.$$

Let $\mathcal{T}_0$ denote the largest ρ_0 coordinates of β in absolute values. Define an active set of strong-signal coordinates, for which we would like to assure recovery, and $\tilde{\rho}_0 \subseteq \mathcal{T}_0 \subset \rho$:

$$\tilde{\rho}_0 = \{j : |\beta_j| > \lambda\sigma\}, \tag{4}$$

We assume standard restricted-eigenvalue conditions and beta-min conditions for support inclusion results.

Assumption 5 (Restricted Eigenvalue Condition $RE(|\rho|, k_0, X)$ (Bickel et al., 2009))**.** *Let X be the data matrix. Define*

$$\frac{1}{\kappa(|\rho|, k_0)} \triangleq \min_{\substack{J_0 \subseteq \{1,\ldots,d\} \\ |J_0| \leq |\rho|}} \min_{\left\|v_{J_0}\right\|_1 \leq k_0 \left\|v_{J_0}\right\|_1} \frac{\|Xv\|_2}{\sqrt{n}\left\|v_{J_0}\right\|_2}.$$

For some integer $1 \leq |\rho| \leq d$ and a number $k_0 > 0$, it holds for all $v \neq 0$,

$$\kappa(|\rho|, k_0)^{-1} > 0,$$

$$\Lambda_{\min}(2|\rho|) := \min_{v \neq 0, \|v\|_0 \leq 2|\rho|} \frac{\|Xv\|_2^2}{n\|v\|_2^2} > 0,$$

$$\Lambda_{\min}(2|\rho|) := \max_{v \neq 0, \|v\|_0 \leq 2|\rho|} \frac{\|Xv\|_2^2}{n\|v\|_2^2} > 0.$$

The restricted eigenvalue condition of Assumption 5 is one of the common assumptions for LASSO. It corresponds to assuming well-conditioning of the matrix under sparse subsets. It also ensures that the behavior policy provides good coverage over relevant features; indeed it characterizes coverage for linear function approximation [Duan et al., 2020].

Assumption 6 (Beta-min condition on strong signals)**.** $\beta_{\min,\tilde{\rho}_0} := \min_{j\in\tilde{\rho}_0} |\beta_j| > \lambda\sigma_r$.

Assumption 6 is a signal-strength condition, that the smallest coordinate of the active set is separated from the threshold defining the active set. This prevents knife-edge situations where a relevant coordinate is not recovered (but is also of irrelevant signal strength). Analogous assumptions are generally required to show support inclusion. Assumption 6 is somewhat milder; instead imposing a stronger version would give correspondingly stronger recovery results.

Under these assumptions, our main result is a prediction error bound on q-function estimation under reward-thresholded lasso, under given rate conditions on threshold and regularization strength of initial lasso.

Theorem 1 (Prediction error bound for reward-thresholded LASSO)**.** *Suppose Assumptions 1 to 6. Suppose Assumption 5,* $\mathrm{RE}\,(\rho_0, 4, X)$ *holds with* $\kappa\,(\rho_0, 4)$*.*

Let β_{init} *be an optimal solution to LASSO*$(\phi, r; \lambda_n)$*, e.g. lasso regression of rewards on features, with* $\lambda_n \geq \frac{\|X\epsilon_\theta\|_\infty}{n}$*. Suppose that for some constants* $\breve{D}_1 \geq D_1$*, and for* $D_0(\Lambda_{\max}, \Lambda_{\min}, |\rho|, \rho_0), D_1(\Lambda_{\max}, \Lambda_{\min}, |\rho|, \rho_0)$ *specified in the appendix, it holds that* $\beta_{\min,\tilde{\rho}_0} \geq D_0\lambda_n\sigma\sqrt{\rho_0} + \breve{D}_1\lambda_n\sigma$*. Choose threshold* $\tau_0 = C\lambda\sigma \geq 2\sqrt{1+a}\lambda\sigma$*, for some constant* $C \geq D_1$*. Let* $\mathcal{I}$ *be the recovered support on* β_{init}*.*

$$\mathcal{I} = \{j : |\beta_{j,init}\,| \geq \tau_0\}\,,\ \textit{where}\ \tau_0 \geq \breve{D}_1\lambda\sigma.$$

Then on $\mathcal{E}_a$*,*

$$\tilde{\rho}_0 \subset \mathcal{I},\ |\mathcal{I}| \leq 2\rho_0,\ \textit{and}\ |\mathcal{I} \cap \mathcal{T}_0^c| \leq \rho_0$$

And, with high probability we have predictive error bounds:

$$\tfrac{1}{n}\|X\hat{\theta} - X\theta^*\|_2^2 \leq 4\tfrac{\sigma_q^2(|\mathcal{I}|(1+468\log(2d))+2(1+2\sqrt{|\mathcal{I}|})}{n}.$$

Given this "fast rate" on the prediction error of the reward-thresholded LASSO, we obtain a bound on the policy error of the fitted-Q-iteration procedure that depends primarily on the *sparsity* (up to constant factors) rather than the potentially *high-dimensional state*. The analysis is standard, given the result we prove above specialized for our method. Note that we did not attempt to optimize problem-independent constants in our analysis.

Before we do so, we show how the thresholded procedure also quantifies an important structural restriction for policy evaluation/optimization: *(approximate) Bellman completeness*, which states that the Bellman operator is approximately closed under the regression function class. Although Proposition 2 establishes that the class of linear functions restricted to the sparse component is Bellman complete, in practice, thresholding noisy estimates may lead to false positives and false negatives. Our previous analysis establishes that these are of controlled magnitude due to the choices of thresholding and regularization parameter. This also implies that the *misspecification bias due to finite-sample estimation* is also vanishing in n at the same rate, stated in the following proposition on approximate instance-dependent Bellman completeness.

Proposition 3 (Bound on Bellman completeness violation under approximate recovery)**.** *With high probability, under* $\mathcal{E}_a$*,*

$$\sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\ \inf_{q_t\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}} \|q_t - \mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2 = O_p(n^{-1}).$$

With these results, we can establish a finite-sample bound on the policy value under Algorithm 2.

Theorem 2. *Suppose Assumptions 1 to 6.*

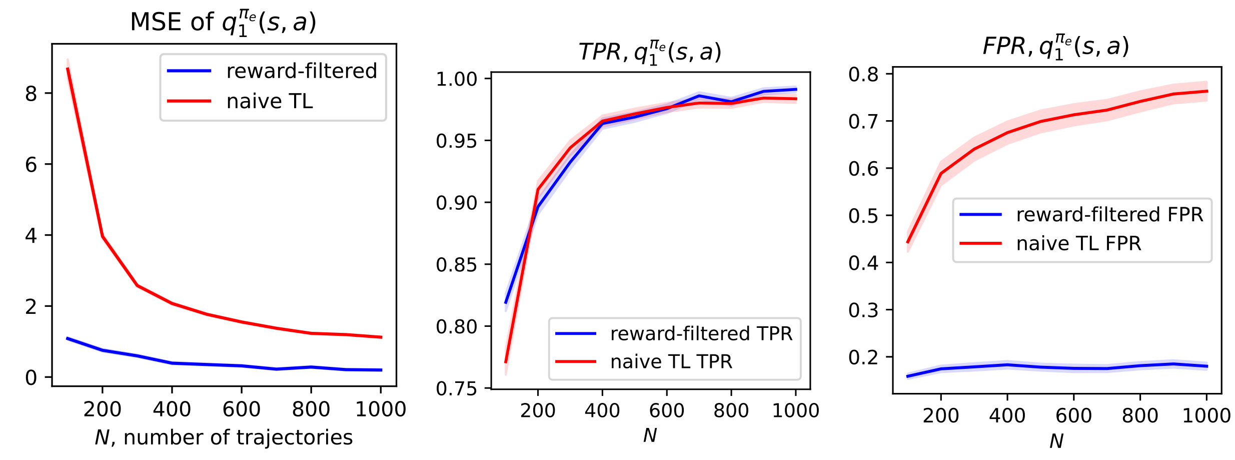
$$V_1^*(s_1) - V_1^\pi(s_1)$$
$$\leq 2T\sqrt{\frac{\Lambda_{\min}\sigma_q^2(938|\rho|\log(2d) + 2(1 + 2\sqrt{|\rho|})}{n}} + o_p(n^{-\frac{1}{2}}).$$

The result follows straightforwardly given our predictive error bound and standard analysis of fitted-Q-iteration. This sample complexity result improves upon prior work since it now depends on the underlying sparsity rather than the full ambient dimension.

7 Experiments

We first consider a simulated setting to validate the method. Our primary comparison is with thresholded LASSO regression for fitted-Q-evaluation. This highlights the benefit of tailoring estimation for the inductive bias. In the data-generating process, we first consider $|\mathcal{S}| = 50, |\rho| = 10$, and $\mathcal{A} = \{0, 1\}$. The reward and states evolve according to

$$r_t(s, a) = \beta^\top\phi_t(s, a) + \epsilon_r,\ \ s_{t+1}(s, a) = M_a s + \epsilon_s.$$



(a) MSE of $\hat{q}_1^{\pi_e}$ under naive thresholding of q-estimation vs. reward-filtered evaluation

(b) True positive rate, recovered support of $\hat{q}_1^{\pi_e}$

(c) False positive rate, recovered support of $\hat{q}_1^{\pi_e}$

Recalling that $M_a = \begin{bmatrix} M_a^{\rho\to\rho} & 0 \\ M_a^{\rho\to\rho_c} & M_a^{\rho_c\to\rho_c} \end{bmatrix}$, we generate the coefficient matrix with independent normal random variables $\sim N(0.2, 1)$. (Note that the nonzero mean helps ensure the beta-min condition). The zero-mean noise terms are normally distributed with standard deviations $\sigma_s = 0.4, \sigma_r = 0.6$. In the estimation, we let $\phi(s, a)$ be a product space over actions, i.e. equivalent to fitting a q function separately for every action.

We first show experiments for policy evaluation in the main text due to space constraints. Fitted-Q-evaluation is similar to fitted-Q-iteration, but replaces the max over q functions with the expectation over actions according to the next time-step's policy. See the appendix for additional experiments for policy optimization specifically. We compare our reward-filtered estimation using Algorithm 2 with naive thresholded lasso, i.e. thresholding lasso-based estimation of q-functions alone in Figures 2a to 2c. (We average the q function over actions; results are similar across actions). The behavior and evaluation policies are both (different) logistic probability models in the state variable, with the coefficient vector given by (different) random draws from the uniform distribution on $[-0.5, 0.5]$. We average over 50 replications from this data generating process and add standard errors, shaded, on the plot. The first plot, Figure 2a, shows the benefits in mean-squared error estimation of the q-function $q_1^{pie}(s, a)$, relative to the oracle q function, which is estimated from a separate dataset of $n = 20000$ trajectories. The reward-filtered method achieves an order of magnitude smaller mean-squared error for small sample sizes, with consistent improvement over thresholded LASSO estimation on the q function alone. Next in Figure 2b we show the true positive rate: both methods perform similarly in including the sparse component the recovered support. But the last plot of Figure 2c shows that the naive thresholded lasso method includes many exogenous variables that are not necessary to recover the optimal policy, while the false positive rate for the reward-filtered method is controlled throughout as a constant fraction of the sparsity. Overall this simple simulation shows the improvements in estimation of the q function (which translate down the line to improvements in decision-value) under this special structure.

Acknowledgments AZ thanks participants at the Simons Program on Causality for inspiring discussions on causal representation learning, even though she still mostly thinks about MDPs. AZ also thanks Dan Malinsky for discussions on graphs and structure at an early stage of this project.

References

K. Ariu, K. Abe, and A. Proutière. Thresholded lasso bandit. In *International Conference on Machine Learning*, pages 878–928. PMLR, 2022.

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 2013.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

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

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

Y. Duan, Z. Jia, and M. Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

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

Y. Efroni, D. Misra, A. Krishnamurthy, A. Agarwal, and J. Langford. Provably filtering exogenous distractors using multistep inverse dynamics. In *International Conference on Learning Representations*, 2021.

Y. Efroni, S. Kakade, A. Krishnamurthy, and C. Zhang. Sparsity in partially controllable linear systems. In *International Conference on Machine Learning*, pages 5851–5860. PMLR, 2022.

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

B. Hao, Y. Duan, T. Lattimore, C. Szepesvári, and M. Wang. Sparse feature selection makes batch reinforcement learning more sample efficient. In *International Conference on Machine Learning*, pages 4063–4073. PMLR, 2021.

D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. pages 9–1, 2012.

A. Lamb, R. Islam, Y. Efroni, A. R. Didolkar, D. Misra, D. J. Foster, L. P. Molu, R. Chari, A. Krishnamurthy, and J. Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Transactions on Machine Learning Research*, 2022.

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

J. N. Lee, G. Tucker, O. Nachum, B. Dai, and E. Brunskill. Oracle inequalities for model selection in offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 28194–28207, 2022.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. 2009.

A. Nedić and D. P. Bertsekas. Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems*, 13(1-2):79–110, 2003.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

M. Seitzer, B. Schölkopf, and G. Martius. Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

S. M. Shortreed and A. Ertefaie. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

S. Van de Geer, P. Bühlmann, and S. Zhou. The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). 2011.

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

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

Z. Wang, X. Xiao, Z. Xu, Y. Zhu, and P. Stone. Causal dynamics learning for task-independent state abstraction. *arXiv preprint arXiv:2206.13452*, 2022b.

A. Zhang, C. Lyle, S. Sodhani, A. Filos, M. Kwiatkowska, J. Pineau, Y. Gal, and D. Precup. Invariant causal prediction for block mdps. In *International Conference on Machine Learning*, pages 11214–11224. PMLR, 2020.

S. Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. *Advances in Neural Information Processing Systems*, 22, 2009.

S. Zhou. Thresholded lasso for high dimensional variable selection and statistical estimation. *arXiv preprint arXiv:1002.1583*, 2010.

Checklist

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, e.g. assumptions about DGP in Sec. 4 and about method in Sec. 6.1]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, sample complexity analysis. We reduce to LASSO, which is classical and has well-described computational complexity elsewhere.]
   (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]
   (b) Complete proofs of all theoretical results. [Yes, in appendix.]
   (c) Clear explanations of any assumptions. [Yes, right after assumptions made.]

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 supplementary material.]
   (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, run on 16gb Macbook with M1 chip]

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]

A Further Discussion

B Further details on method

Choosing the penalty level in practice A data driven suggestion of [Belloni and Chernozhukov, 2013] is to choose

$$\lambda = \frac{c'\hat{\sigma}\Lambda(1-\alpha \mid X)}{n}$$

where $\Lambda(1-\alpha \mid X)$ is the $(1-\alpha)$ quantile of $n\|S/\sigma\|_\infty$. They also suggest to choose a data-driven upper bound for $\hat{\sigma}_0$ the sample deviation of y_i, compute LASSO, and then set $\hat{\sigma}^2 = \hat{q}(\hat{\beta})$.

C Proofs

C.1 Proofs of characterization

Proof of Proposition 1. The proof follows by induction. We first show the base case, for $t = T$. Recall that we take the convention that $r_T = q_T = 0$, so that $q_T(s,a) = r_T(s,a)$. Therefore since $r_T(s,a) = r_T(\tilde{s},a)$ for $s_T, \tilde{s}_T$ such that $s_T^\rho = \tilde{s}_T^\rho$, we also have that for

$$\pi_T^*(s) \in \arg\max_{a\in\mathcal{A}} r(s,a), \qquad \tilde{\pi}_T^*(\tilde{s}) \in \arg\max_{a\in\mathcal{A}} r(\tilde{s},a),$$

and when $s_T^\rho = \tilde{s}_T^\rho$, $\pi_T^*(s_T) = \tilde{\pi}_T^*(\tilde{s}_T)$. Therefore $q_T^*(s_T,a) = q_T^*(\tilde{s}_T,a)$ when $s^\rho = \tilde{s}^\rho$. Next we show the inductive step. The inductive hypothesis is that

$$q_{t+1}^*(s_{t+1},a) = q_{t+1}^*(s_{t+1}^\rho,a) = q_{t+1}^*(\tilde{s}_{t+1},a), \forall a \in \mathcal{A}, \text{ and } \pi_{t+1}^*(s_{t+1}) = \tilde{\pi}_{t+1}^*(\tilde{s}_{t+1}) \text{ when } s_{t+1}^\rho = \tilde{s}_{t+1}^\rho.$$

Then for

$$\begin{aligned}
\pi_t^*(s) &\in \arg\max_{a\in\mathcal{A}}\{r_t(s,a) + \gamma\mathbb{E}[q_{t+1}^*(s_{t+1},\pi_{t+1}^*(s_{t+1})) \mid s,a]\} \\
\tilde{\pi}_t^*(\tilde{s}) &\in \arg\max_{a\in\mathcal{A}}\{r_t(\tilde{s},a) + \gamma\mathbb{E}[q_{t+1}^*(\tilde{s}_{t+1},\pi_{t+1}^*(s_{t+1})) \mid \tilde{s},a]\}
\end{aligned}$$

we have that

$$\begin{aligned}
q_t^*(s_t,a) &= r_t(s_t,a) + \gamma\mathbb{E}[q_{t+1}^*(s_{t+1},\pi_{t+1}^*(s_{t+1})) \mid s_t,a] \\
&= r_t(s_t,a) + \gamma\mathbb{E}[q_{t+1}^*(s_{t+1}^\rho,\pi_{t+1}^*(s_{t+1}^\rho)) \mid s_t,a] && \text{(induction hypothesis)} \\
&= r_t(s_t,a) + \gamma\mathbb{E}[q_{t+1}^*(s_{t+1}^\rho,\pi_{t+1}^*(s_{t+1}^\rho)) \mid s_t^\rho,a] && \text{(Assumption 3)} \\
&= r_t(\tilde{s}_t,a) + \gamma\mathbb{E}[q_{t+1}^*(\tilde{s}_{t+1},\pi_{t+1}^*(\tilde{s}_{t+1})) \mid \tilde{s}_t^\rho,a] && (s_t^\rho = \tilde{s}_t^\rho) \\
&= q_t^*(\tilde{s}_t,a)
\end{aligned}$$

when $s_t^\rho = \tilde{s}_t^\rho$.

Therefore, when $s_t^\rho = \tilde{s}_t^\rho$,

$$\pi_t^*(s_t) = \tilde{\pi}_t^*(\tilde{s}_t).$$

This completes the induction. □

Proof of Proposition 2. Let $\pi^*(s_{t+1}) \in \arg\max_{a\in\mathcal{A}} \breve{q}(s_{t+1},a)$. Note that when $\breve{q} \in \breve{\Pi}_{t+1}$, the optimal action remains the same for states that differ only outside of the sparse support: $\pi^*(s_{t+1}) = \pi^*(\tilde{s}_{t+1})$ when $s_{t+1}^\rho = \tilde{s}_{t+1}^\rho$.

Therefore for any $\breve{q} \in \breve{\Pi}_{t+1}$,

$$\begin{aligned}
\mathcal{T}^*\breve{q} &= \mathbb{E}_{s_{t+1}^\rho}\left[\mathbb{E}_{s_{t+1}^{\rho c}}\left[\breve{q}_{t+1}(s_{t+1},a^*(s_{t+1})) \mid s_{t+1}^\rho,s,a\right] \mid s,a\right] \\
&= \mathbb{E}_{s_{t+1}^\rho}\left[q_{t+1}(s_{t+1}^\rho,a^*(s_{t+1}^\rho)) \mid s,a\right] \\
&= \mathbb{E}_{s_{t+1}^\rho}\left[q_{t+1}(s_{t+1}^\rho,a^*(s_{t+1}^\rho)) \mid s^\rho,a\right] && \text{by Assumption 3}
\end{aligned}$$

where the second-to-last equality holds since $\breve{q}_{t+1}(s_{t+1}, a^*(s_{t+1})) = \breve{q}_{t+1}(\tilde{s}_{t+1}, a^*(s_{t+1}))$ when $s^\rho_{t+1} = \tilde{s}^\rho_{t+1}$, for any $\breve{q}_{t+1} \in \breve{q}_{t+1}$.

Next we show that under Assumptions 1 and 3, $\mathbb{E}[r_t(s,a) + \mathcal{T}^* \breve{q}_{t+1} \mid s, a]$ is linear and is representable by a function $\breve{q} \in \amalg_t$. Under linear rewards, $r_t(s,a) = \theta^*_t \phi_\rho(s,a)$ for some θ^*_t that is ρ-sparse. And, under linear transitions, $\mathbb{E}_{s^\rho_{t+1}}[\mathcal{T}^* \breve{q}_{t+1} \mid s_t, a_t] = \phi_\rho^\top \mu_\rho^{*,\top} \breve{q}^*_{t+1}$ where μ^*_ρ is the $\rho-$marginalized linear transition map. Hence

$$\mathbb{E}[r_t(s,a) + \mathcal{T}^* \breve{q}_{t+1} \mid s, a] = \underbrace{(\theta^*_t + \breve{q}^{*,\top}_{t+1} \mu^*_\rho)}_{w^*_\rho} \phi_\rho(s,a)$$

□

C.2 Intermediate results

We first study the parameter error of ordinary least squares under a missing set of covariates. We let $\mathcal{I}$ denote the subset of covariates, for example that returned by thresholded lasso. We first consider the case when $\mathcal{I}$ is a given subset containing the true support. The next theorem is a more complex extension, specialized to our reward-thresholded q-estimation setting, of a result about estimation under omitted variables of [Zhou, 2010, 2009]. The key structure allowing us to link thresholded lasso of reward to prediction error of estimated q functions is the shared covariance structure. Theorem 3 is the main technical contribution of our work.

Theorem 3 (Prediction error bounds of $\mathcal{I}$-restricted ordinary least squares of the Bellman residual)**.** *Suppose Assumptions 1 to 5. Let $\mathcal{D} = \{1, \ldots, d\} \setminus \mathcal{I}$ and $\rho_\mathcal{D} = \mathcal{D} \bigcap \rho$ (e.g. the set of false negatives of support recovery). Suppose $|\rho \bigcup \rho_\mathcal{D}| \leq 2|\rho|$ and that $\rho \subseteq \mathcal{I}$. Suppose that $\lambda \geq \|\frac{X^\top \epsilon_{r+q}}{\sqrt{n}}\|_\infty$. Consider $\mathcal{I}-$restricted ordinary least squares regression of sparse q. In the following, we omit the time index for brevity. Then:*

$$\|\hat{\theta}_\mathcal{I} - \theta_\mathcal{I}\|_2 \leq \frac{\sqrt{|\mathcal{I}|}}{\Lambda_{\min}(|\mathcal{I}|)} \lambda. + \|\theta_\mathcal{D}\|_2$$

$$\frac{1}{n}\|X\hat{\theta} - X\theta^*\|_2^2 \leq 4\frac{\sigma_q^2(|\mathcal{I}| + 2\sqrt{|\mathcal{I}|\log(1/\delta)} + \log(1/\delta))}{n} + \max\left(36\frac{|\mathcal{I}|\lambda^2}{\kappa}, 162\frac{\sigma_\theta^2 |\mathcal{I}| \log(d/|\mathcal{I}|)}{n}\right)$$

Proof of Theorem 3. Let θ be the full ordinary least squares solution for the q estimation, $\hat{\theta}_\mathcal{I}, \theta^*_\mathcal{I}$ be the estiamted and true restricted OLS solution computed on $\mathcal{I}$, respectively, and θ^* the true (sparse) solution.

$$\begin{aligned}
\hat{\theta}_\mathcal{I} &= (X_\mathcal{I}^\top X_\mathcal{I})^{-1} X_\mathcal{I}^\top Y(q) = (X_\mathcal{I}^\top X_\mathcal{I})^{-1} X_\mathcal{I}^\top Y(q) \\
&= (X_\mathcal{I}^\top X_\mathcal{I})^{-1} X_\mathcal{I}^\top (X_\mathcal{I}^\top \theta^*_\mathcal{I} + \epsilon) && \text{(sparse rewards and asn. about sparse } Q \text{ function)} \\
&= \theta^*_\mathcal{I} + (X_\mathcal{I}^\top X_\mathcal{I})^{-1} X_\mathcal{I}^\top \epsilon_{r+\gamma q}
\end{aligned}$$

Hence,

$$\|\hat{\theta}_\mathcal{I} - \theta^*_\mathcal{I}\|_2 \leq \|(X_\mathcal{I}^\top X_\mathcal{I})^{-1} X_\mathcal{I}^\top \epsilon_{r+\gamma q}\|_2.$$

We bound the second term as follows:

$$\|(X_\mathcal{I}^\top X_\mathcal{I})^{-1} X_\mathcal{I}^\top \epsilon_{r+\gamma q}\|_2 \leq \left\|\left(\frac{X_\mathcal{I}^\top X_\mathcal{I}}{n}\right)^{-1}\right\|_2 \left\|\frac{X_\mathcal{I}^\top \epsilon_{r+\gamma q}}{n}\right\|_2 \leq \frac{\sqrt{|\mathcal{I}|}}{\Lambda_{\min} s(|\mathcal{I}|)} \lambda,$$

yielding that

$$\|\hat{\theta}_\mathcal{I} - \theta^*_\mathcal{I}\|_2 \leq \frac{\sqrt{|\mathcal{I}|}}{\Lambda_{\min}(|\mathcal{I}|)} \lambda.$$

The result follows since $\|\hat{\theta}_\mathcal{I} - \theta^*\|_2^2 \leq 2\|\hat{\theta}_\mathcal{I} - \theta^*_\mathcal{I}\|_2^2 + 2\|\theta^*_\mathcal{I} - \theta^*\|_2^2$.

Next we bound the prediction error,

$$\frac{1}{n}\|X\hat{\theta} - X\theta^*\|_2^2 \leq \frac{1}{n}\|X\hat{\theta} - X\theta^*_\mathcal{I}\|_2^2 + \frac{1}{n}\|X\theta^*_\mathcal{I} - X\theta^*\|_2^2.$$

We will decompose relative to $\hat{\beta}$, a thresholded lasso regression on rewards r alone but also restricted to $\mathcal{I}$. Note that by the elementary bound $(a-b)b \leq (a-b)^2 + b^2$:

$$\begin{aligned}\frac{1}{n}\|X\hat{\theta} - X\theta^*_{\mathcal{I}}\|_2^2 &\leq \frac{2}{n}\|X\hat{\theta} - X\theta^*_{\mathcal{I}} - (X\hat{\beta} - X\beta^*_{\mathcal{I}})\|_2^2 + \frac{2}{n}\|X\hat{\beta} - X\beta^*_{\mathcal{I}}\|_2^2 \\ &\coloneqq T_1 + T_2.\end{aligned}$$

First we bound $T_1 \coloneqq \frac{2}{n}\|X\hat{\theta} - X\theta^*_{\mathcal{I}} - (X\hat{\beta} - X\beta^*_{\mathcal{I}})\|_2^2$. Observe that it is equivalently the prediction error when regressing the next-stage q function alone, i.e. $y_t - r_t$, on the $\mathcal{I}$-restricted features, since $(\hat{\theta} - \hat{\beta}) = (X_{\mathcal{I}}^\top X_{\mathcal{I}})^{-1} X_{\mathcal{I}}^\top \{\gamma v(s')\}$. Then

$$\frac{2}{n}\left\|X(\hat{\theta} - \hat{\beta}) - X(\theta^*_{\mathcal{I}} - \beta^*_{\mathcal{I}})\right\|_2^2 = \frac{2}{n}\left\|X\left\{(X_I^\top X_I)^{-1}X_I^\top(\gamma v(s_{t+1})) - (\theta^*_{\mathcal{I}} - \beta^*_{\mathcal{I}})\right\}\right\|_2^2,$$

where the last term can be identified as the noise term in $(V(s_{t+1})) - (\theta^*_{\mathcal{I}} - \beta^*_{\mathcal{I}})) \approx \epsilon_q$ under the linear MDP assumption. By the sparsity properties of $\hat{\theta}, \hat{\beta}$ (they are both restricted to $\mathcal{I}$):

$$\begin{aligned}\frac{2}{n}\left\|X(\hat{\theta} - \hat{\beta}) - X(\theta^*_{\mathcal{I}} - \beta^*_{\mathcal{I}})\right\|_2^2 &= \frac{2}{n}\left\|X_{\mathcal{I}}(\hat{\theta} - \hat{\beta}) - X_{\mathcal{I}}(\theta^*_{\mathcal{I}} - \beta^*_{\mathcal{I}})\right\|_2^2 && \text{(by two-step procedure and realizability)} \\ &\leq \frac{\sigma_q^2(2|\mathcal{I}| + 2\sqrt{2|\mathcal{I}|\log(1/\delta)} + 2\log(1/\delta))}{n}. && \text{(by Lemma 1)}\end{aligned}$$

Next we bound $T_2 \coloneqq \frac{2}{n}\|X\hat{\beta} - X\beta^*_{\mathcal{I}}\|_2^2$. Let β^λ denote the initial LASSO solution in the thresholded lasso $\hat{\beta}$.

By optimality of $\hat{\beta}$,

$$\begin{aligned}\frac{2}{n}\|X\hat{\beta} - X\beta^*_{\mathcal{I}}\|_2^2 &\leq \frac{2}{n}\|X\beta^\lambda - X\beta^*_{\mathcal{I}}\|_2^2 \\ &\leq \frac{4}{n}\|X\beta^\lambda - X\beta^*\|_2^2 + \frac{4}{n}\|X\beta^* - X\beta^*_{\mathcal{I}}\|_2^2\end{aligned}$$

The first of these is bounded via standard analysis of prediction error in LASSO, and the second by a maximal inequality as previously.

By the penalized formulation:

$$\begin{aligned}\frac{1}{2n}\left\|X\beta^\lambda - X\beta^*\right\|_2^2 &\leq \frac{\lambda}{2}\left\|\beta^\lambda - \beta^*\right\|_1 + \lambda\left(\|\beta^*\|_1 - \|\beta^\lambda\|_1\right) \\ &\leq \frac{\lambda}{2}\left\|\beta^\lambda_{\mathcal{I}} - \beta^*_{\mathcal{I}}\right\|_1 + \lambda\left\|\beta^\lambda_{\mathcal{I}_c}\right\|_1 + \lambda\left(\|\beta^*\|_1 - \|\beta^\lambda\|_1\right) \\ &\leq \frac{\lambda}{2}\left\|\beta^\lambda_{\mathcal{I}} - \beta^*_{\mathcal{I}}\right\|_1 + \lambda\left\|\beta^\lambda_{\mathcal{I}_c}\right\|_1 + \lambda\left(\left\|\beta^*_{\mathcal{I}} - \beta^\lambda_{\mathcal{I}}\right\|_1 - \left\|\beta^\lambda_{\mathcal{I}_c}\right\|_1\right) \\ &= \frac{3\lambda}{2}\left\|\beta^\lambda_{\mathcal{I}} - \beta^*_{\mathcal{I}}\right\|_1 - \frac{\lambda}{2}\left\|\beta^\lambda_{\mathcal{I}_c}\right\|_1,\end{aligned}$$

The above, with the restricted eigenvalue condition of Assumption 5, implies that

$$\frac{3\lambda}{2}\left\|\beta^\lambda_{\mathcal{I}} - \beta^*_{\mathcal{I}}\right\|_1 - \frac{\lambda}{2}\left\|\beta^\lambda_{\mathcal{I}_c}\right\|_1 \geq \frac{1}{2n}\|X\beta^\lambda - X\beta^*\|_2^2 \geq \kappa\|\beta^\lambda - \beta^*\|_2^2 \tag{5}$$

Therefore, by properties of the ℓ_1 and ℓ_2 norm:

$$\frac{1}{2n}\|X\beta^\lambda - X\beta^*\|_2^2 \leq \frac{3\lambda}{2}\|\beta^\lambda_{\mathcal{I}} - \beta^*_{\mathcal{I}}\|_1 \leq \frac{3\lambda\sqrt{|\mathcal{I}|}}{2}\|\beta^\lambda_{\mathcal{I}} - \beta^*_{\mathcal{I}}\|_2$$

Then applying the restricted eigenvalue condition of Assumption 5 to the last term of the above, we obtain that

$$\frac{1}{2n}\left\|X\beta^\lambda - X\beta^*\right\|_2^2 \leq \frac{3\lambda\sqrt{|\rho|}\left\|X\beta^\lambda - X\beta^*\right\|_2}{\sqrt{n\kappa}}.$$

Rearranging, this gives the bound

$$\frac{4}{n}\left\|X\beta^{\lambda} - X\beta^{*}\right\|_2^2 \leq 144\frac{|\mathcal{I}|\lambda^2}{\kappa}.$$

Finally, we can bound $\frac{2}{n}\|X\beta^*_{\mathcal{I}} - X\beta^*\|_2^2$ via a maximal inequality over the ℓ_0 norm ball of radius $2\rho_0$ since earlier we showed that $|\mathcal{I}| \leq 2\rho_0$. Applying the maximal inequality of Lemma 2 gives

$$\frac{4}{n}\|X\beta^*_{\mathcal{I}} - X\beta^*\|_2^2 \leq 324\frac{\sigma_\theta^2|\mathcal{I}|\log(d/|\mathcal{I}|)}{n}.$$

□

C.3 Proofs of main results for method

Proof of Theorem 1. Because the support is recovered from a thresholded LASSO on the rewards, the support inclusion result is a consequence of [Zhou, 2010, Thm. 6.3], although analogous results essentially hold under stronger beta-min conditions (i.e, on the support ρ and correspondingly stronger support inclusion conditions). Namely, it gives that, suppose for some constants $\breve{D}_1 \geq D_1$, and for D_0, D_1 such that: For $K := \kappa(\rho_0, 6)$, $b_0 \geq 2$,

$$\begin{aligned}
&D_0 = \max\left\{D, K\sqrt{2}\left(2\sqrt{\Lambda_{\max}(|\rho| - \rho_0)} + 3b_0K\right)\right\} \\
&\quad \text{where } D = (\sqrt{2}+1)\frac{\sqrt{\Lambda_{\max}(|\rho| - \rho_0)}}{\sqrt{\Lambda_{\min}(2\rho_0)}} + \frac{\theta_{\rho_0,2\rho_0}\Lambda_{\max}(|\rho| - \rho_0)}{\Lambda_{\min}(2\rho_0)} \text{ and} \\
&D_1 = 2\Lambda_{\max}(|\rho| - \rho_0)/b_0 + 9K^2b_0/2,
\end{aligned}$$

it holds that, for $\breve{D}_1 \geq D_1$,

$$\beta_{\min,A_0} \geq D_0\lambda\sigma\sqrt{\rho_0} + \breve{D}_1\lambda\sigma, \text{ where } \lambda := \sqrt{2\log p/n}.$$

Choose a thresholding parameter τ_0 and set

$$\mathcal{I} = \{j : |\beta_{j,\text{init}}| \geq \tau_0\}, \text{ where } \tau_0 \geq \breve{D}_1\lambda\sigma.$$

Then on $\mathcal{E}_a$,

$$\tilde{\rho}_0 \subset \mathcal{I}, |\mathcal{I} \cap \mathcal{T}_0^c| \leq \rho_0, |\mathcal{I}| \leq 2\rho_0, \tag{6}$$

$$\|\beta_{\mathcal{D}}\|_2^2 \leq (\rho_0 - a_0)\lambda^2\sigma^2. \tag{7}$$

This yields the first statement about support recovery.

For prediction error, we then apply Theorem 1 and this yields the result. □

Proof of Proposition 3. True β^* is ρ-sparse but the worst case situation is if $\rho \setminus \tilde{\rho}_0 \not\subseteq \mathcal{I}$, i.e. the low-signal coefficients are not returned by the thresholding algorithm. On the other hand, they are assuredly of magnitude $\leq |\lambda\sigma|$ and hence ought to lead to less violation of the completeness condition. Let $\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}$ denote the set of linear coefficients with support on $\|\mathcal{I}\|_0 \leq 2\rho_0$ such that it does not contain the low signal variables $\rho \setminus \tilde{\rho}_0$, and

$$\sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}} \inf_{q_t\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}} \|q_t - \mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2 \leq \epsilon.$$

The infimum over q_t is equivalent to a further-restricted ℓ_0 norm regression problem.

$$\begin{aligned}
&\sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}} \inf_{q_t\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}} \|q_t - \mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2 \\
&= \inf_{q_t\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}} \sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}} \|q_t - \mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2
\end{aligned}$$

and

$$\begin{aligned}&\sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\{\|q_t-\mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2+\|q_t-\mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2\}\\ &\leq \sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\|q_t-\mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2+\sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\|\mathcal{T}_t^\star q_{t+1}^*-\mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2\end{aligned}$$

Then

$$\sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\|\mathcal{T}_t^\star q_{t+1}^*-\mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2\leq(\|(q_{t+1})_{\mathcal{I}\setminus\tilde{\rho}_0}\|_1+\|(q_{t+1})_{\rho\setminus\tilde{\rho}_0}\|_1)^2\leq(2s\tau_0+\sqrt{2s}\|\hat{\beta}-\beta\|_2+s\lambda\sigma)^2$$

That is, false positives are of low signal strength (by the algorithm, and by prediction error bound) while false negatives not in the active set are also of low signal strength. The threshold and signal strength definitions tend to 0 at a rate overall depending on λ. Therefore, using a (loose) bound that $(a+b)^2\leq 2a^2+2b^2$,

$$\sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\|\mathcal{T}_t^\star q_{t+1}^*-\mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2\leq(2s(\tau_0+\lambda\sigma)+\sqrt{2s}\|\hat{\beta}-\beta\|_2)^2\leq 16s^2(\tau_0^2+\lambda^2\sigma^2)+4s\|\hat{\beta}-\beta\|_2^2$$

Next we bound:

$$\inf_{q_t\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\ \sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\|q_t-\mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2$$

The outer minimization is simply least-squares regression over a further restricted ℓ_0 norm ball. Consider $\tilde{\mathcal{Q}}$ such that $\tilde{\mathcal{Q}}=\{q\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}\colon q_{\tilde{\rho}_0}>0, q_{\mathcal{I}\setminus\tilde{\rho}_0}=0\}$, and note that $\tilde{\mathcal{Q}}\subseteq\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}$.

$$\inf_{q_t\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\ \sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\|q_t-\mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2\leq\inf_{q_t\in\tilde{\mathcal{Q}}}\ \sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\|q_t-\mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2$$

The worst-case error is incurred when $q_{t+1,\rho\setminus\tilde{\rho}_0}>0$; these are the low-signal variables not guaranteed to be recovered by the algorithm. Then for $q'\in\mathcal{Q}_{\setminus\tilde{\rho}_0}:=\{q\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}\colon q_{\rho\setminus\tilde{\rho}_0}>0\}$, and we have that

$$\leq\inf_{q_t\in\tilde{\mathcal{Q}}}\ \sup_{q_{t+1}\in\mathcal{Q}_{\setminus\tilde{\rho}_0}}\|q_t-\mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2,$$

where the leading order dependence is described by Theorem 3's analysis of least-squares regression on a restricted covariate set: $\tilde{\mathcal{Q}}$ omits the low-signal variables $\rho\setminus\tilde{\rho}_0$. Therefore, by Theorem 3, w.h.p. under $\mathcal{E}_a$ and assumptions on λ in Theorem 3,

$$\sup_{q_{t+1}\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\ \inf_{q_t\in\mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0\not\subseteq\mathcal{I}}}\|q_t-\mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2=O_p(n^{-1}).$$

□

C.4 Technical results

We list standard technical results from other works that we use without proof.

C.4.1 Concentration

Lemma 1 (Theorem 1 of [Hsu et al., 2012], random design prediction bound for linear regression.)**.** *Define* $\widehat{\Sigma}:=\widehat{\mathbb{E}}[x\otimes x]=\frac{1}{n}\sum_{i=1}^n x_i\otimes x_i$*. Suppose outcomes are* σ_{noise}*-subgaussian and "bounded statistical leverage", then there exists a finite* $\rho_{2,cov}\geq 1$ *such that almost surely:*

$$\frac{\|\Sigma^{-1/2}X\|}{\sqrt{d}}=\frac{\|\Sigma^{-1/2}X\|}{\sqrt{\mathbb{E}[\|\Sigma^{-1/2}X\|^2]}}\leq\rho_{2,\mathrm{cov}}$$

If $n > n_{2,\delta}$, then with probability at least $1 - 2\delta$, we have that the matrix error $\left\|\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}\right\| \leq K_{2,\delta,n} \leq 5$; and the excess loss satisfies:

$$\|\hat{w}_{\text{ols}} - w\|_{\Sigma}^2 \leq K_{2,\delta,n} \cdot \frac{\sigma_{noise}^2 \cdot (d + 2\sqrt{d\log(1/\delta)} + 2\log(1/\delta))}{n}$$

Lemma 2 (Prediction error bounds via maximal inequalities over an ℓ_0 ball, Theorem 4 of [Raskutti et al., 2011] .)**.** *For any covariate matrix X, with probability greater than $1 - \exp(-10s\log(d/s))$ the minimax prediction risk is upper bounded as*

$$\min_{\widehat{w}} \max_{w^* \in \mathbb{B}_0(|\mathcal{I}|)} \frac{1}{n}\left\|X\left(\widehat{w} - w^*\right)\right\|_2^2 \leq 81\frac{\sigma^2|\mathcal{I}|\log(d/|\mathcal{I}|)}{n},$$

where $\mathbb{B}_0(|\mathcal{I}|)$ is the ℓ_0 norm ball of radius $|\mathcal{I}|$.

C.4.2 Analysis of fitted-Q-evaluation

Definition 3 (Bellman error)**.** *Under data distribution μ_t, define the Bellman error of function $q = (q_0, \ldots, q_{T-1})$ as: $\mathcal{E}(q) = \frac{1}{T}\sum_{t=0}^{T-1}\|q_t - \mathcal{T}_t^* q_{t+1}\|_{\mu_t}^2$*

Lemma 3 (Bellman error to value suboptimality)**.** *Under Assumption 5, for any $q \in \mathcal{Q}$, we have that, for π the policy that is greedy with respect to q, $V_1^*(s_1) - V_1^\pi(s_1) \leq 2T\sqrt{C \cdot \mathcal{E}(q^\pi)}$.*

Proof of Theorem 2. Under Lemma 3, it suffices to bound the Bellman error, $\frac{1}{T}\sum_{t=0}^{T-1}\|q_t - \mathcal{T}_t^* q_{t+1}\|_{\mu_t}^2$. We start with one timestep. Let $\ell(f,g) = (f-g)^2$ be the squared error. The Bellman error satisfies that $\|\hat{q}_h - \mathcal{T}_h^\star \hat{q}_{h+1}\|_{\mu_h}^2$ and can be lower bounded as follows:

$$\begin{aligned}\|\hat{q}_h - \mathcal{T}_h^\star \hat{q}_{h+1}\|_{\mu_h}^2 &= \mathbb{E}_{\mu_h}[\ell(\hat{q}_h, \hat{q}_{h+1})] - \mathbb{E}_{\mu_h}[\ell(q_h^\dagger, \hat{q}_{h+1})] + \|q_h^\dagger - \mathcal{T}_h^\star \hat{q}_{h+1}\|_{\mu_h}^2 \\ &\leq \mathbb{E}_{\mu_h}[\ell(\hat{q}_h, \hat{q}_{h+1})] + \epsilon \qquad \text{(by Proposition 3 on apx. Bellman completeness)}\end{aligned}$$

where ϵ is the parameter for approximate Bellman completeness, such that

$$\sup_{q_{t+1} \in \mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0 \not\subseteq \mathcal{I}}} \inf_{q_t \in \mathcal{Q}_{\mathcal{I},\rho\setminus\tilde{\rho}_0 \not\subseteq \mathcal{I}}} \|q_t - \mathcal{T}_t^\star q_{t+1}\|_{\mu_t}^2 \leq \epsilon.$$

By Proposition 2 we have that $\epsilon = O_p(n^{-1})$.

The prediction error bound of Theorem 1 bounds $\mathbb{E}_{\mu_h}[\ell(\hat{q}_h, \hat{q}_{h+1})]$ so we have that

$$V_1^*(s_1) - V_1^\pi(s_1) \leq 2T\sqrt{\frac{\Lambda_{\min}\sigma_q^2(2|\rho|(1 + 468\log(2d)) + 2(1 + 2\sqrt{|\rho|})}{n}}.$$

□

D Alternative model: endogenous/exogenous decomposition of Dietterich et al. [2018]

We discuss a related, but different model: a sparse reward variant of the endogenous-exogenous variable decomposition of Dietterich et al. [2018]. The main difference is that the exogenous components instead can affect the endogenous components, as opposed to the other way around in our model, where endogenous components affect exogenous components. We include the illustration in Figure 3.

A natural question is whether our methods can handle this setting as well, especially since Dietterich et al. [2018] shows that the optimal policy is sparse in the endogenous MDP alone. Our exact characterization in this paper used the conditional independence restriction of Assumption 3, which does not hold in the exo-endo MDP since exogenous variables can affect next-timestep endogenous variables. On the other hand,

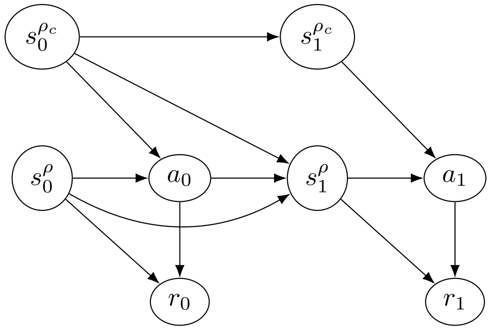


Figure 3: "Exogenous/endogenous MDP" of [Dietterich et al., 2018].

that the optimal policy is sparse in the endogenous MDP alone implies that the corresponding *advantage functions*, i.e. $\Delta_{a_0}(s,a) = q(s,a) - q(s,a_0)$ do in fact satisfy the conditional independence restriction of Assumption 3.

Hence, under the additional restriction of reward sparsity where exogenous variables do not affect reward, we can extend methods in this paper to thresholded-LASSO based on estimating reward *contrast* functions and hence advantage functions. To sketch this extension, note that we can run CATE estimation at the final timestep and then simply redefine Bellman targets to be differences of q-functions over actions.

This additional assumption of reward sparsity is required: in the original paper of [Dietterich et al., 2018], rewards are additively decomposable but there can be direct effect of exogenous variables on the reward.

E Experiments

In the data-generating process, we first consider $|\mathcal{S}| = 50, |\rho| = 10$, and $\mathcal{A} = \{0,1\}$. The reward and states evolve according to

$$r_t(s,a) = \beta^\top \phi_t(s,a) + \epsilon_r, \;\; s_{t+1}(s,a) = M_a s + \epsilon_s.$$

Recalling that $M_a = \begin{bmatrix} M_a^{\rho\to\rho} & 0 \\ M_a^{\rho\to\rho_c} & M_a^{\rho_c\to\rho_c} \end{bmatrix}$, we generate the coefficient matrix with independent normal random variables $\sim N(0.2,1)$. (Note that the nonzero mean helps ensure the beta-min condition). The zero-mean noise terms are normally distributed with standard deviations $\sigma_s = 0.4, \sigma_r = 0.6$. In the estimation, we let $\phi(s,a)$ be a product space over actions, i.e. equivalent to fitting a q function separately for every action.

We first show experiments for policy evaluation in the main text due to space constraints. Fitted-Q-evaluation is similar to fitted-Q-iteration, but replaces the max over q functions with the expectation over actions according to the next time-step's policy. See the appendix for additional experiments for policy optimization specifically. We compare our reward-filtered estimation using Algorithm 2 with naive thresholded lasso, i.e. thresholding lasso-based estimation of q-functions alone in Figures 2a to 2c. (We average the q function over actions; results are similar across actions). The behavior and evaluation policies are both (different) logistic probability models in the state variable, with the coefficient vector given by (different) random draws from the uniform distribution on $[-0.5, 0.5]$. We average over 50 replications from this data generating process and add standard errors, shaded, on the plot. The first plot, Figure 2a, shows the benefits in mean-squared error estimation of the q-function $q_1^{pi_e}(s,a)$, relative to the oracle q function, which is estimated from a separate dataset of $n = 20000$ trajectories. The reward-filtered method achieves an order of magnitude smaller mean-squared error for small sample sizes, with consistent improvement over thresholded LASSO estimation on the q function alone. Next in Figure 2b we show the true positive rate: both methods perform similarly in including the sparse component the recovered support. But the last plot of Figure 2c shows that the naive thresholded lasso method includes many exogenous variables that are not necessary to recover the optimal policy, while the false positive rate for the reward-filtered method is controlled throughout as a constant fraction of the sparsity. Overall this simple simulation shows the improvements in estimation of the

q function (which translate down the line to improvements in decision-value) under this special structure.