Sparse Feature Selection Makes Batch Reinforcement Learning More Sample Efficient

Botao Hao [1] **Yaqi Duan** [2] **Tor Lattimore** [1] **Csaba Szepesvári** [1 3] **Mengdi Wang** [2 1]

Abstract

This paper provides a statistical analysis of high-dimensional batch reinforcement learning (RL) using sparse linear function approximation. When there is a large number of candidate features, our result sheds light on the fact that sparsity-aware methods can make batch RL more sample efficient. We first consider the off-policy evaluation problem. To evaluate a new target policy, we analyze a Lasso fitted Q-evaluation method and establish a finite-sample error bound that has no polynomial dependence on the ambient dimension. To reduce the Lasso bias, we further propose a post model-selection estimator that applies fitted Q-evaluation to the features selected via group Lasso. Under an additional signal strength assumption, we derive a sharper instance-dependent error bound that depends on a divergence function measuring the distribution mismatch between the data distribution and occupancy measure of the target policy. Further, we study the Lasso fitted Q-iteration for batch policy optimization and establish a finite-sample error bound depending on the ratio between the number of relevant features and restricted minimal eigenvalue of the data's covariance. In the end, we complement the results with minimax lower bounds for batch-data policy evaluation/optimization that nearly match our upper bounds. The results suggest that having well-conditioned data is crucial for sparse batch policy learning.

1. Introduction

We consider batch reinforcement learning (RL), where the problem is to evaluate a target policy or to learn a good policy based on a given dataset (Szepesvári, 2010; Lange et al., 2012; Levine et al., 2020). While in online RL the central question is how to sequentially interact with the environment to balance exploration and exploitation, in batch RL the dataset is given a priori and the focus is typically on learning a near-optimal policy or evaluating a given target policy.

To handle RL systems with large or even infinite state spaces, we focus on the use of linear function approximation (Bellman et al., 1963; Schweitzer and Seidmann, 1985; Bertsekas and Tsitsiklis, 1996), that is, using a weighted linear combination of available features (aka basis functions) to represent high-dimensional transition/value functions. Results from the supervised learning literature show that the sample size needed to get accurate policy evaluations or near-optimal policies must scale at least linearly with d, the number of features (e.g., Example 15.14 of Wainwright, 2019).

We leverage the idea of sparse approximation and focus on situations when a smaller number of *unknown* relevant features is sufficient for solving the RL problem. Sparse regression has proved to be a powerful method for high-dimensional statistical learning with limited data (Tibshirani, 1996; Chen et al., 2001; Bunea et al., 2007; Bickel et al., 2009; Rish and Grabarnik, 2014), and we will borrow techniques from the sparse learning literature to improve the sample efficiency of batch RL, an idea with a considerable history in RL as witnessed by our literature review that follows below.

Contribution. We make several contributions:

- First, we consider the problem of off-policy evaluation (OPE). To promote sparse solutions, we iteratively fits state-action value functions using linear regression with ℓ_1-regularization which we call this procedure as *Lasso fitted Q-evaluation* (Algorithm 1). We estab-

[1]Deepmind [2]Princeton University [3]University of Alberta. Correspondence to: Botao Hao <haobotao000@gmail.com>, Mengdi Wang <mengdiw@princeton.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

lish the first finite-sample error bound for Lasso fitted Q-evaluation that depends linearly on the number of relevant features and the restricted minimal eigenvalue. Importantly, the bound has no polynomial dependence on d. This appears to be the first theoretical bound for sparse batch policy evaluation.

- Second, to reduce the Lasso bias, we propose an improved *post model-selection estimator* (Algorithm 2) that applies fitted Q-evaluation with a smaller feature set that is selected using group Lasso. Under an additional separability assumption, we derive a sharper and nearly minimax-optimal error bound that is instance-dependent. The error bound is determined by a divergence function measuring the distribution mismatch, *restricted over the reduced feature space*, between the data distribution and the occupancy distribution of the target policy. This divergence defined over the reduced feature space is significantly smaller than its counterpart over the full d-dimensional space. In other words, *sparse feature selection reduces the distribution mismatch*. We also provide a nearly-matching lower bound, and these two results sharply characterize the statistical limits of sparse off-policy evaluation.

- Third, we extend our analysis to the batch policy optimization problem.We analyze the *Lasso fitted Q-iteration* (Algorithm 3) and show that the ℓ_∞-norm of policy error depends linearly on the ratio between the number of relevant features and the restricted minimal eigenvalue of the dataset's covariance matrix. Finally, we establish a minimax lower bound for sparse batch policy learning and show that the lower bound also depends on the aforementioned ratio. This is the first lower bound result, to the authors' best knowledge, demonstrating the critical role played by the minimal eigenvalue of the dataset's covariance matrix and the construction is highly non-trivial. The upper and lower bounds validate the belief that well-conditioned data is crucial for sample-efficient policy learning.

2. Preliminaries

2.1. Problem definition

A finite, infinite-horizon discounted Markov decision process (DMDP) can be described by the tuple $M = (\mathcal{X}, \mathcal{A}, P, r, \gamma)$. Here, $\mathcal{X}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $P : \mathcal{X} \times \mathcal{A} \to \Delta_{\mathcal{X}}$ is the transition probability function, $r : \mathcal{X} \times \mathcal{A} \to [0, 1]$ is the reward function and $\gamma \in (0, 1)$ is the so-called discount factor. In this paper, for the sake of simplicity, we stick to finite DMDPs. However, our results can be extended to more general cases with routine work.

We define a (stationary) policy π as a $\mathcal{X} \to \Delta_{\mathcal{A}}$ map, mapping states to distributions over actions. A policy, a distribution ξ_0 over $\mathcal{X}$ and a DMDP M together give rise to a probability measure $\mathbb{P}^\pi$ over the set of infinitely long state-action histories: for a history of form $x_1, a_1, r_1, x_2, a_2, r_2, \ldots$, $\mathbb{P}^\pi(x_1) = \xi_0(x_1)$ and for $t = 1, 2, \ldots$, $\mathbb{P}^\pi(a_t|x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t) = \pi(a_t|x_t)$ and $\mathbb{P}^\pi(x_{t+1}|x_1, a_1, \ldots, x_{t-1}, a_{t-1}, x_t, a_t) = P(x_{t+1}|x_t, a_t)$. We denote by $\mathbb{E}^\pi$ the corresponding expectation operator. The *value* of policy π given the initial state distribution ξ_0 is $v^\pi_{\xi_0} := \mathbb{E}^\pi[\sum_{t=0}^\infty \gamma^t r(x_t, a_t)]$, where $\mathbb{E}^\pi$ depends on ξ_0 but just this dependence is not shown. We define $v^\pi : \mathcal{X} \to \mathbb{R}$, the *value function* of policy π, by letting $v^\pi(x)$ denote the value of π when it is started from state x.

A nonstationary policy is a sequence of maps from histories to probability distributions over actions and such a policy similarly induces a probability measure $\mathbb{P}^\pi$ and an underlying expectation operator $\mathbb{E}^\pi$. An *optimal policy* π^* maximizes the value from every state: $v^{\pi^*}(x) = \max_\pi v^\pi(x)$ for any $x \in \mathcal{X}$ where the maximum is over all policies, including nonstationary ones. As is well known, for finite DMDPs, a stationary optimal policy always exist (e.g., Szepesvári, 2010). The value function shared by optimal policies is called the *optimal value function* and is denoted by v^*.

We consider the following learning and optimization problems. The learner knows the state space $\mathcal{X}$ and action space $\mathcal{A}$. The reward function r is given in the form of a black box, which the learner can use to evaluate $r(x, a)$ for any pair of $(x, a) \in \mathcal{X} \times \mathcal{A}$. The only unknown is the transition probability function P. The learner is given a random dataset $\mathcal{D} = \{(x_n, a_n, x'_n)\}_{n=1}^N$ generated by using a (possibly nonstationary and unknown) *behavior policy* $\bar{\pi}$ in the DMDP M starting from some initial distribution which may be different from ξ_0. We study two fundamental batch RL tasks:

- *Off-policy policy evaluation:* given $\mathcal{D}$ and black box access to a *target policy* π, ξ_0 and r, estimate the value, $v^\pi_{\xi_0}$, of π;

- *Batch policy optimization:* given $\mathcal{D}$ and black box access to r, find an optimal policy.

Bellman operators. By slightly abusing the notation, we will view the transition probability kernel P as a left linear operator, mapping from $\mathbb{R}^{\mathcal{X}}$ to $\mathbb{R}^{\mathcal{X}\times\mathcal{A}}$:

$$(Pv)(x,a) := \sum_{x'} P(x'|x,a)v(x').$$

The *Bellman optimality operator* $\mathcal{T} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ is defined as

$$[\mathcal{T}v](x) = \max_{a\in\mathcal{A}}(r(x,a) + \gamma Pv(x,a)),\ \forall x \in \mathcal{X}.$$

The function $v^* \in \mathbb{R}^{\mathcal{X}}$ is the unique solution to the Bellman optimality equation $v = \mathcal{T}v$. The state-action value function of a policy π is defined as

$$Q^\pi(x,a) = r(x,a) + \gamma Pv^\pi(x,a).$$

We also introduce the Bellman operator $\mathcal{T}_\pi : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ for policy π as:

$$[\mathcal{T}_\pi v](x) = \sum_{a\in\mathcal{A}} \pi(a|x)\left(r(x,a) + \gamma Pv(x,a)\right)\ ,\ \forall x \in \mathcal{X}.$$

The function v^π is the unique solution to the Bellman equation $v^\pi = \mathcal{T}_\pi v^\pi$.

Linear function approximation. Let $\phi : \mathcal{X}\times\mathcal{A} \to \mathbb{R}^d$ be a feature map which assigns to each state-action pair a d-dimensional feature vector. A feature map combined with a parameter vector $w \in \mathbb{R}^d$ gives rise to the linear function $g_w : \mathcal{X}\times\mathcal{A} \to \mathbb{R}$ defined by $g_w(x,a) = \phi(x,a)^\top w$, $(x,a) \in \mathcal{X}\times\mathcal{A}$, and the linear function class $\mathcal{G}_\phi = \{g_w : w \in \mathbb{R}^d\}$. When little a priori information is available on how to choose the features, agnostic choices often lead to dimensions which can be as large as the number of samples n. However, effective learning with many more features than the sample-size is possible when only $s \ll d$ features are relevant. Thus, we further consider a sparse linear function class $\mathcal{G}_{\phi,\mathcal{K},s} = \{g_w : w \in \mathbb{R}^d, w_{\mathcal{K}^c} = 0, |\mathcal{K}| \le s\}$.

Fix a feature map ϕ. For $f : \mathcal{X}\times\mathcal{A} \to \mathbb{R}$, let $f^\pi : \mathcal{X} \to \mathbb{R}$ be defined by $f^\pi(x) = \sum_a \pi(a|x)f(x,a)$. We present a representation condition for $\mathcal{G}_{\phi,\mathcal{K},s}$ that ensures sample-efficient estimation.

Assumption 2.1 (Policy completeness). For some feature map ϕ, $\mathcal{G}_{\phi,\mathcal{K},s}$ is closed under the operators $f \mapsto Pf^\pi$ for any policy π of the DMDP.

Policy completeness requires function class $\mathcal{G}_{\phi,\mathcal{K},s}$ can well capture Bellman operator. It is crucial for estimation consistency of fitted-Q-iteration type algorithm (Le et al., 2019; Duan and Wang, 2020) and implies the realizability condition regarding Q^π. Recently, Wang et al. (2020) has derived a lower bound to show that Q^π realizability condition only is not sufficient for a sample-efficient estimation of OPE with linear function approximation. More discussions regarding those conditions in batch policy learning could refer Chen and Jiang (2019).

Our second assumption concerns the dataset:

Assumption 2.2. The dataset $\mathcal{D}$ consists of $N = KL$ samples from K independent episodes $\tau_1, \ldots, \tau_K$. Each episode τ_k has L consecutive transitions generated by some unknown behavior policy $\bar{\pi}_k$ giving rise to a sample path $\tau_k = (x_0^{(k)}, a_0^{(k)}, x_0^{(k)'}, \ldots, x_{L-1}^{(k)}, a_{L-1}^{(k)}, x_{L-1}^{(k)'})$.

3. Sparsity-Aware Off-Policy Evaluation

In this section we consider OPE problem, i.e., to estimate the value of a target policy π from logged experiences $\mathcal{D}$ generated using unknown behavior policies. We propose two sparsity-aware algorithms to approximate state-action functions using sparse parameters.

3.1. Lasso fitted Q-evaluation

We propose a straightforward modification of the fitted Q-evaluation method (Le et al., 2019; Duan and Wang, 2020) to account for sparsity. For $t = 1, 2, \ldots, T-1$, the algorithm produces $\widehat{w}_{t+1}$ using Lasso-regularized regression. To make the errors of different steps independent, the dataset $\mathcal{D}$ is split into T nonoverlapping folds $\mathcal{D}_1, \ldots, \mathcal{D}_T$, so that all folds have $R = K/T$ episodes in them and only data from fold t is used in step t. To define the algorithm, it is useful to introduce $Q_w = r + \gamma g_w$ where $w \in \mathbb{R}^d$. For $a < b$, we also define the operator $\Pi_{[a,b]} : \mathbb{R} \to [a,b]$ that projects its input to $[a,b]$, i.e., $\Pi_{[a,b]}(x) = \max(\min(x,b),a)$. The pseudocode is given as Algorithm 1. In the last step, $m = N$ samples are used to produce the final output to guarantee that the error introduced by the Monte-Carlo averaging is negligible compared to the rest.

Remark 3.1. The last step Monte Carlo averaging is only for numerical integration, where the samples are newly drawn inside the algorithm (independent from batch data), so there is no bias here. We set $m = N$ to simplify the theory but it could be much larger than N for a more accurate approximation.

3.2. Post model-selection fitted Q-evaluation

Sparse regularization is known to induce a small bias in regression. However, this bias could get compounded through the iterative procedure of Algorithm 1. To avoid such bias and improve the accuracy, we aim to identify the set of rel-

Algorithm 1 Lasso fitted Q-evaluation

```
Input: Initial distribution $\xi_0$, target policy $\pi$, $T$ folds of dataset $\{\mathcal{D}_t\}_{t=1}^T$ of overall size $N$, regularization parameter $\lambda_1 > 0$, $m := N$, $\widehat{w}_0 = 0 \in \mathbb{R}^d$.
for $t = 1, 2, \ldots, T$ do
    Calculate regression targets for $(x_i, a_i, x_i') \in \mathcal{D}_t$:
        $y_i = \sum_a \pi(a|x_i') Q_{\widehat{w}_{t-1}}(x_i', a)$ .
    Fit $\widehat{w}_t$ through sparse linear regression $\widehat{w}_t = \operatorname{argmin}_w \Big\{ \lambda_1 \|w\|_1 + \frac{1}{|\mathcal{D}_t|} \sum_{(x_i, a_i, x_i') \in \mathcal{D}_t} (\Pi_{[0,1/(1-\gamma)]} y_i - \phi(x_i, a_i)^\top w)^2 \Big\}$ .
end for
Output: $\widehat{v}_{\widehat{w}_T}^{\pi} = \frac{1}{m} \sum_{u=1}^m \Pi_{[0,1/(1-\gamma)]} (Q_{\widehat{w}_T}(\widetilde{x}_u, \widetilde{a}_u))$ with $\widetilde{x}_u \sim \xi_0$, $\widetilde{a}_u \sim \pi(\cdot|\widetilde{x}_u)$.
```

evant features $\mathcal{K}$ before evaluating the policy based on the following proposition.

Proposition 3.2. Under Assumption 2.1, there exists a $d \times d$ matrix K^π such that $\forall\ (x, a) \in \mathcal{X} \times \mathcal{A}$

$$\mathbb{E}_{x' \sim P(\cdot|x,a)}[\phi^\pi(x')^\top \mid x, a] = \phi(x, a)^\top K^\pi, \tag{3.1}$$

where $\phi^\pi(x) = \sum_a \pi(a|x)\phi(x, a)$ and all but s rows of K^π are identically zero.

Thus we propose to estimate the set of relevant features $\mathcal{K}$ using group lasso (Yuan and Lin, 2006). Once the relevant feature set is identified, any regular policy evaluation method can be used over the learned feature set $\widehat{\mathcal{K}}$. In Algorithm 2, for the ease of comparability with the previous method, we consider vanilla fitted Q-evaluation.

Remark 3.3. One may wonder whether it is necessary to refit the iterative regression and why not simply use the estimated $\widehat{K}^\pi$ to get a plug-in estimator. This is because refitting typically performs strictly better than direct regularized learning and has less bias, as long as the feature selection succeeds (Belloni et al., 2013).

4. Performance Bounds For Sparse Off-Policy Evaluation

Algorithm 2 Post model-selection fitted Q-evaluation

```
Input: initial distribution $\xi_0$, target policy $\pi$, dataset $\mathcal{D}$ of size $N$, $m := N$, number of iterations $T$, $\lambda_2, \lambda_3 > 0$.
Estimate $\widehat{K}^\pi$ through: $\widehat{K}^\pi = \operatorname{argmin}_{K \in \mathbb{R}^{d \times d}} \Big\{ \lambda_2 \sum_{j=1}^d \|K_{j\cdot}\|_2 + \frac{1}{Nd} \sum_{n=1}^N \big\|\phi^\pi(x_n')^\top - \phi(x_n, a_n)^\top K\big\|_2^2 \Big\}$.   (3.2)
Find $\widehat{\mathcal{K}} = \{j \in [d] : \|\widehat{K}_{j\cdot}^\pi\|_2 \neq 0\}$, the estimated set of relevant features. Here, $\widehat{K}_{j\cdot}^\pi$ refers to the $j$th row of $\widehat{K}^\pi$.
Initialize: $\widehat{w}_0 = 0 \in \mathbb{R}^{|\widehat{\mathcal{K}}|}$.
for $t = 1, 2, \ldots, T$ do
    Calculate regression targets for $n \in [N]$:
        $y_n = r(x_n, a_n) + \gamma \sum_a \pi(a|x_n') [\phi(x_n', a)_{\widehat{\mathcal{K}}}]^\top \widehat{w}_{t-1}$ .
    Update $\widehat{w}_t$ through $\widehat{w}_{t+1} = \operatorname{argmin}_{w \in \mathbb{R}^{|\widehat{\mathcal{K}}|}} \Big\{ \frac{1}{N} \sum_{n=1}^N (y_n - [\phi(x_n, a_n)_{\widehat{\mathcal{K}}}]^\top w)^2 + \lambda_3 \|w\|_2^2 \Big\}$ .
end for
Output: $\widehat{v}_{\widehat{w}_T}^{\pi} = \frac{1}{m} \sum_{u=1}^m \Pi_{[0,1/(1-\gamma)]} (Q_{\widehat{w}_T}(\widetilde{x}_u, \widetilde{a}_u))$ with $\widetilde{x}_u \sim \xi_0$, $\widetilde{a}_u \sim \pi(\cdot|\widetilde{x}_u)$.
```

We study the finite-sample estimation error of Algorithms 1, 2. All the technical proofs are deferred to the Appendix. Let Σ be the expected uncentered covariance matrix of the batch data, given by

$$\Sigma := \mathbb{E}\left[\frac{1}{L} \sum_{h=0}^{L-1} \phi(x_h^{(1)}, a_h^{(1)}) \phi(x_h^{(1)}, a_h^{(1)})^\top\right], \tag{4.1}$$

where L is the length of one episode.

Remark 4.1. Our analysis applies to multiple behavior policies and our algorithms do not require the knowledge of the behavior policy. For example, if we have M behavior policies $\{\pi_1, \ldots, \pi_M\}$ and one policy is selected following distribution p_π at the beginning of each episode to collect data, we could define Σ similarly but the expectation will take also w.r.t the randomness with p_π.

We need a notion of restricted eigenvalue that is common in high-dimensional statistics (Bickel et al., 2009; Bühlmann and Van De Geer, 2011).

Definition 4.2 **(Restricted eigenvalue).** Given a positive semi-definite matrix $Z \in \mathbb{R}^{d \times d}$ and integer $s \geq 1$, define the restricted minimum eigenvalue of Z as $C_{\min}(Z,s) :=$

$$\min_{\mathcal{S}\subset[d],|\mathcal{S}|\le s}\ \min_{\boldsymbol{\beta}\in\mathbb{R}^d}\left\{\frac{\langle\boldsymbol{\beta},Z\boldsymbol{\beta}\rangle}{\|\boldsymbol{\beta}_{\mathcal{S}}\|_2^2}:\|\boldsymbol{\beta}_{\mathcal{S}^c}\|_1\le 3\|\boldsymbol{\beta}_{\mathcal{S}}\|_1\right\}.$$

4.1. Finite-sample error bounds of Algorithm 1

We first provide a statistical error bound for the Lasso fitted Q-evaluation (Algorithm 1).

Theorem 4.3 (OPE error bound for Algorithm 1). Suppose Assumptions 2.1, 2.2 hold and $C_{\min}(\Sigma,s) > 0$. Let Algorithm 1 take N samples satisfying $N \gtrsim s^2\log(d/\delta)L(1-\gamma)^{-1}/C_{\min}(\Sigma,s)$. Set the number of iterations $T = \Theta(\log(N/(1-\gamma))/(1-\gamma))$ and $\lambda_1 = (1-\gamma)^{-1}\sqrt{T\log(2d/\delta)/N}$. Then, with probability at least $1-2\delta$,

$$|\widehat{v}^{\pi}_{\widehat{w}_T} - v^{\pi}| \lesssim \frac{1}{C_{\min}(\Sigma,s)}\sqrt{\frac{s^2\log(d/\delta)}{N(1-\gamma)^5}}. \tag{4.2}$$

Theorem 4.3 shows that the OPE error depends linearly on $s/C_{\min}(\Sigma,s)$. For comparison, when the sparsity is not considered, Duan and Wang (2020) proved the error bound (using our notations) of the form

$$|\widehat{v}^{\pi} - v^{\pi}| \lesssim \sqrt{d/(C_{\min}(\Sigma,d)N(1-\gamma)^4)}.$$

From the Definition 4.2, $C_{\min}(\Sigma,d) < C_{\min}(\Sigma,s)$. Comparing the two results (setting $\delta = 1/N$), we expect the new error bound to be significantly tighter , i.e., $C_{\min}(\Sigma,d)\frac{s^2\log(dN)}{1-\gamma} \ll C^2_{\min}(\Sigma,s)d$, when there is a high level of sparsity ($s \ll d$).

Remark 4.4. The restricted eigenvalue $C_{\min}(\Sigma,s)$ quantifies how well the behavior policy covers the state-action feature space. We need $C_{\min}(\Sigma,s) > 0$ meaning that the data is well-conditioned or the behavior policy provides good coverage over relevant features. In the tabular case, where $\phi(x,a)$ is a basis vector in $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{A}|}$, then $C_{\min} = \min_{x,a}\mu^{\pi}(x,a) > 0$ implies that the behavior policy visits all state-action pairs with positive probability, where $\mu^{\pi}(x,a)$ is the visitation distribution under policy π and initial distribution ξ_0: $\mu^{\pi}(x,a) = \frac{1}{L}\sum_{h=0}^{L-1}\mathbb{E}^{\pi}\left[\mathbb{I}((x_h,a_h) = (x,a))\right]$.

To ensure the success of policy evaluation/optimization with linear function approximation, similar assumptions regarding Σ in RL literature also appear in Abbasi-Yadkori et al. (2019), Duan and Wang (2020), Lazic et al. (2020). In addition, Agarwal et al. (2020) showed how to design a policy that can collect data that satisfies this condition.

4.2. Finite-sample error bounds of Algorithm 2

Next, we give a result for the post-selection model estimator for OPE (Algorithm 2). We will show that this algorithm provides a more accurate estimate under the additional condition that every relevant feature plays a nontrivial role.

Assumption 4.5 (Minimal signal strength). For some given $\delta > 0$, the *minimum signal strength* satisfies

$$\min_{j\in\mathcal{K}}\|K^{\pi}_{j\cdot}\|_2/\sqrt{d} \ge \frac{64\sqrt{2}s}{C_{\min}(\Sigma,s)}\sqrt{\frac{2\log(2d^2/\delta)}{N}},$$

where $K^{\pi}_{j\cdot}$ is the jth row of K^{π} defined in Eq. (3.1).

Then we provide a critical lemma showing that the group lasso step in Algorithm 2 is guaranteed to identify a sufficiently sparse feature set including all the relevant features with high probability.

Lemma 4.6 (Feature screening). Suppose Assumptions 2.1, 4.5 hold and $C_{\min}(\Sigma,s) > 0$. Set the regularization parameter $\lambda_2 = 4\sqrt{2\log(2d^2/\delta)/(Nd)}$ for some $\delta > 0$ and let the sample size satisfy $N \gtrsim Ls^2\log(d/\delta)/C^2_{\min}(\Sigma,s)$. Then with probability at least $1-\delta$, the size of learned relevant feature set $\widehat{\mathcal{K}}$ satisfies $|\widehat{\mathcal{K}}| \lesssim s$ and $\widehat{\mathcal{K}} \supseteq \mathcal{K}$ where $\mathcal{K}$ is the true relevant feature set of M.

The detailed proof of Lemma 4.6 is deferred to Appendix B.2. This lemma implies the model group lasso selected is sufficiently sparse and can identify all the important features with high probability. Thus, we have $|\widehat{\mathcal{K}}| \asymp s$.

Now we analyze the policy evaluation error of Algorithm 2. According to Cramer-Rao lower bound for tabular OPE (Jiang and Li, 2016) and the minimax lower bound for OPE with linear function approximation (Duan and Wang, 2020), we expect the optimal OPE error to depend on the *distribution mismatch* between the target policy and the behavior policy that generated the data. To define the notion of distribution mismatch, we first need the notion of occupancy measures:

Definition 4.7 (Occupancy measures). Let $\bar{\mu}$ be the expected occupancy measure of observations $\{(x_n,a_n)\}_{n=1}^N$: $\bar{\mu}(x,a) = \sum_{n=1}^N\mathbb{P}(x_n = x, a_n = a)/N$ and μ^{π} be the discounted occupancy distribution of (x_h,a_h) under policy π and initial distribution ξ_0: $\mu^{\pi}(x,a) = (1-\gamma)\mathbb{E}^{\pi}[\sum_{h=0}^{\infty}\gamma^h\mathbb{1}(x_h = x, a_h = a)], \forall\, x,a$.

Inspired by Theorem 5 of Duan and Wang (2020), we will measure the distribution mismatch using *restricted chi-square divergences* between $\bar{\mu}$ and μ^{π}.

Definition 4.8 (Restricted chi-square divergence). Let $\mathcal{G}$ be a set of real-valued functions over $\mathcal{X}$ and let p_1 and p_2 be probability distributions over $\mathcal{X}$. We define the *$\mathcal{G}$-restricted chi-square divergence* (or $\chi^2_{\mathcal{G}}$-divergence) between p_1 and p_2 as

$$\chi^2_{\mathcal{G}}(p_1, p_2) := \sup_{f\in\mathcal{G}} \frac{\mathbb{E}_{p_1}[f(x)]^2}{\mathbb{E}_{p_2}[f(x)^2]} - 1.$$

By using the feature screening Lemma 4.6, and a similar analysis as by Duan and Wang (2020), we obtain the following instance-dependent error bound for sparse off-policy evaluation.

Theorem 4.9 (Instance-dependent error bound for sparse OPE). Suppose Assumptions 2.1, 2.2 and 4.5 hold and $C_{\min}(\Sigma, s) > 0$. Let $\delta \in (0, 1)$ and assume that Algorithm 2 is fed with N samples satisfying $N \gtrsim L\log(d^2/\delta)s^2/C^2_{\min}(\Sigma, s) + \gamma^2 L\log(s/\delta)s/(1-\gamma)^2$. Set $\lambda_2 = 4\sqrt{2\log(2d^2/\delta)/Nd}, \lambda_3 = \lambda_{\min}(\Sigma)\log(12|\widehat{\mathcal{K}}|/\delta)L|\widehat{\mathcal{K}}|$. Letting the number of iterations $T \to \infty$, the following holds with probability at least $1 - 3\delta$,

$$|\widehat{v}^{\pi}_{\widehat{w}} - v^{\pi}| \lesssim \sqrt{1 + \chi^2_{\mathcal{G}(\widehat{\mathcal{K}})}(\mu^{\pi}, \bar{\mu})}\sqrt{\frac{\log(1/\delta)}{N(1-\gamma)^4}}, \qquad (4.3)$$

where $\bar{\mu}$ is the data generating distribution, $\mathcal{G}(\widehat{\mathcal{K}})$ is the reduced feature space.

Remark 4.10 (Reduced distribution mismatch via sparse feature selection.). The OPE error bound of Theorem 4.9 depends on the statistics $\chi^2_{\mathcal{G}(\widehat{\mathcal{K}})}(\mu^{\pi}, \bar{\mu})$ that quantifies the distribution mismatch between data and the target policy. This result implies the uncertainty for evaluating a new policy from batch data crucially and jointly depends on the two distributions as well as the function class used for fitting. When $\widehat{\mathcal{K}}$ is a small subset of $[d]$, we have $\chi^2_{\mathcal{G}(\widehat{\mathcal{K}})} \ll \chi^2_{\mathcal{G}([d])}$. Therefore our instance-dependent error bound is expected to be significantly smaller than its counterpart that does not exploit sparsity.

4.3. Minimax lower bound for sparse OPE

To complete the picture, we provide a minimax lower bound of off-policy evaluation using sparse linear function approximation. The proof is an adaptation of the respective lower bound proof for linear case (Theorem 3 in Duan and Wang (2020)). It implies the bound in Theorem 4.9 is nearly minimax-optimal.

Theorem 4.11. There exists a DMDP instance with feature map $\phi(\cdot,\cdot) \in \mathbb{R}^d$ satisfying $\|\phi(x, a)\|_\infty \le 1$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ and Assumption 2.1 such that if $N \gtrsim sL(1-\gamma)^{-1}$, then

$$\inf_{\widehat{v}^{\pi}} \sup_{\phi, M\in\mathcal{M}_{\phi,s}(\mathcal{X},\mathcal{A},\gamma)} \mathbb{P}_M\Big(|\widehat{v}^{\pi}(\mathcal{D}) - v^{\pi}| \gtrsim \frac{1}{(1-\gamma)^2}\sqrt{1 + \chi^2_{\mathcal{G}(\mathcal{K})}(\mu^{\pi}, \bar{\mu})}\sqrt{\frac{1}{N}}\Big) \ge \frac{1}{6},$$

where $\phi \in (\mathbb{R}^d)^{\mathcal{X}\times\mathcal{A}}$, $\mathbb{P}_M$ is the probability measure under the DMDP instance M and $\widehat{v}^{\pi}(\cdot)$ sweeps through all algorithms that estimate values based on data $\mathcal{D}$.

Remark 4.12. It is worth to mention that in Theorem 4.11, the distribution mismatch term $1 + \chi^2_{\mathcal{G}(\mathcal{K})}(\mu^{\pi}, \bar{\mu})$ may also contain a $1-\gamma$ term *in the worse case*. Thus the lower bound of sparse OPE also has a $\sqrt{1/(1-\gamma)^3}$ dependency in the worse case that matches the result for the lower bound of sparse batch policy optimization in Theorem 5.2. This matches the result in Yin and Wang (2020) (Theorem 3.1&Remark 3.3) since they define n as the number of episodes while we define N as the number of total sample pairs.

5. Sparsity-Aware Batch Policy Optimization

We extend our analysis to batch policy learning problem with sparse linear function approximation. Consider the Lasso fitted Q-iteration (see Algorithm 3) that has been studied in Calandriello et al. (2014) as a special case of an algorithm for sparse multi-task RL. It resembles Algorithm 1 except for that it calculates the regression target with an additional "max" operation.

Algorithm 3 Lasso-regularized fitted Q-iteration

```
Input: $T$ folds of dataset $\{\mathcal{D}_t\}_{t=1}^T$, regularization parameter $\lambda_1$, $\widehat{w}_0 = 0 \in \mathbb{R}^d$.
Repeat:
for $t = 1, 2, \dots, T$ do
    Calculate regression targets: for $(x_i, a_i, x'_i) \in \mathcal{D}_t$, $y_i = \max_{a\in\mathcal{A}} Q_{\widehat{w}_{t-1}}(x'_i, a)$.
    Based on $\{(\Pi_{[0,1/(1-\gamma)]}y_i, \phi(x_i, a_i))\}_{(x_i,a_i,x'_i)\in\mathcal{D}_t}$, fit $\widehat{w}_t$ through Lasso as in Algorithm 1.
end for
Output: policy $\widehat{\pi}_T(\cdot|x) = \max_{a\in\mathcal{A}} Q_{\widehat{w}_T}(x, a), \forall x \in \mathcal{X}$.
```

The next theorem proves the approximate optimality of the learned policy using Lasso fitted Q-iteration.

Theorem 5.1. Suppose Assumptions 2.1-2.2 holds and $C_{\min}(\Sigma, s) > 0$. Let $N \gtrsim s^2L(1-\gamma)^{-1}/C_{\min}(\Sigma, s)$. Let Algorithm 3 take $T = \Theta(\log(N/(1-\gamma))/(1-\gamma))$ and

$\lambda_1 = (1-\gamma)^{-1}\sqrt{T\log(2d/\delta)/N}$. Then, with probability at least $1-\delta$,

$$\left\|v^{\widehat{\pi}_T} - v^*\right\|_\infty \lesssim \frac{s}{C_{\min}(\Sigma, s)}\sqrt{\frac{\log(d/\delta)}{N(1-\gamma)^7}}. \tag{5.1}$$

Theorem 5.1 suggests that the sample size needed to get a good policy depends mainly on the number of relevant features s, instead of the large ambient dimension d, provided that the data is well-conditioned. This result is not surprising: Calandriello et al. (2014) gave a similar upper bound for sparse FQI for the setting of generative model. Le et al. (2019) provided a generalization theory for policy evaluation/learning with a general function class and their error bound depends on the VC-dimension of the class, but it requires a stronger coefficient concentration condition.

5.1. Minimax lower bound for policy optimization

In the end, we study the fundamental limits of sparse batch policy learning. We establish an information-theoretic minimax lower bound that nearly match the aforementioned upper bound.

Theorem 5.2. Let $\gamma \geq \frac{2}{3}$ and $\widehat{\pi}$ denote an algorithm that maps dataset $\mathcal{D}$ to a policy $\widehat{\pi}(\mathcal{D})$. If $N \gtrsim sL(1-\gamma)^{-1}$, then for any $\widehat{\pi}$, there always exists a DMDP instance M with feature map $\phi(\cdot,\cdot) \in \mathbb{R}^d$ satisfying $\|\phi(x,a)\|_\infty \leq 1$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ and Assumption 2.1, such that

$$\mathbb{P}_M\left(v^*_{\xi_0} - v^{\widehat{\pi}(\mathcal{D})}_{\xi_0} \gtrsim \sqrt{\frac{1+\chi^2_{\mathcal{G}(\mathcal{K})}(\mu^*, \bar{\mu})}{N(1-\gamma)^3}}\right) \geq \frac{1}{6}, \tag{5.2}$$

where μ^* is the discounted state-action occupancy measure of π^*. In addition, we have

$$\mathbb{P}_M\left(v^*_{\xi_0} - v^{\widehat{\pi}(\mathcal{D})}_{\xi_0} \gtrsim \sqrt{\frac{s/C_{\min}(\Sigma, s)}{N(1-\gamma)^3}}\right) \geq \frac{1}{6}. \tag{5.3}$$

Theorems 5.1, 5.2 show that the statistical error of batch policy learning is fundamentally determined by the ratio $s/C_{\min}(\Sigma, s)$. Note that there remains a gap $\sqrt{s/C_{\min}(\Sigma, s)}$ between Theorems 5.1 and 5.2, due to the nature of Lasso regression.

Remark 5.3. Minimax sample complexity lower bound for solving MDP has been studied in the setting with a generative model that allows querying any (s,a) for independent samples. Azar et al. (2013) constructed a hard instance of tabular MDP and, by reducing policy optimization to a testing a Bernoulli distribution, proved a lower bound $SA/(1-\gamma)^3$ which is known to be sharp. Yang and Wang (2019) extended the construction to linear MDP and show that the sample complexity lower bound is $d/(1-\gamma)^3$ under a generative model. There also exists matching upper bound in the same setting.

Our Theorem 5.2 applies to the setting of batch episodic data where are highly dependent. Due to this major difference, we have to use a more intricate proof based on likelihood test to establish a minimax lower bound. Further, Theorem 5.2 characterizes for the first time that the lower bound depends on the minimal eigenvalue of the data's population covariance.

Remark 5.4. Earlier results such as those of (Munos and Szepesvári, 2008; Antos et al., 2008; Le et al., 2019) require stronger forms of concentration condition that the state-action occupancy measure (or a ratio involving this measure) is entrywisely bounded across all policies. Such entrywise bound can be very large if the state-action space $\mathcal{X}$ is large. In contrast, our results only require that the data's covariance Σ is well-conditioned on restricted supports, which is a much weaker assumption.

Further, one can use the empirical minimal eigenvalue to get a rough error estimate. Theorem 5.2 further validates that the minimal eigenvalue indeed determines the statistical limit of batch policy optimization. The result is the first of its kind to our best knowledge.

6. Related Work

Off-policy evaluation (OPE). OPE often serves the starting point of batch RL. A direct approach was to fit value function from data using approximate dynamic programming, e.g., the policy evaluation analog of fitted Q-iteration (Ernst et al., 2005; Munos and Szepesvári, 2008; Le et al., 2019) or least square policy iteration (Lagoudakis and Parr, 2003). Another popular class of OPE methods used importance sampling to get unbiased value estimate of a new policy (Precup et al., 2000) and improved by doubly-robust technique to reduce the variance (Jiang and Li, 2016; Thomas and Brunskill, 2016). To alleviate the curse of horizon (Li et al., 2015; Jiang and Li, 2016; Yin and Wang, 2020), marginalized importance sampling was suggested by estimating state marginal importance ratio without reweighting the entire trajectory (Hallak and Mannor, 2017; Liu et al., 2018; Xie et al., 2019). In general, estimating marginalized importance ratio could be sample-expensive and even intractable. Recently, practical duality-inspired methods were developed for estimating this ratio using function approximation (Nachum et al., 2019; Uehara and Jiang, 2019; Zhang et al., 2020a;b; Yang et al., 2020).

On the theoretical side, Uehara and Jiang (2019); Yin and Wang (2020); Kallus and Uehara (2020) established asymptotic optimality and efficiency for OPE in the tabular setting. Duan and Wang (2020) showed that fitted Q-evaluation with linear function approximation is minimax optimal and provided matching upper and lower bounds that depend on a distribution mismatch term. Another closely related work was by Le et al. (2019) who studied batch policy evaluation and optimization with more general function approximation. They showed the complexity of batch RL depends on the complexity of the function class, assuming a "concentration coefficient" condition (Munos and Szepesvári, 2008) that the state-action visitation density is bounded entrywisely across policies. More recently, Uehara and Jiang (2019) provided theoretical investigations into OPE using general function approximators for marginalized importance weights and value functions but did not show the statistical optimality.

Sparse learning in RL. The use of feature selection by regularization in RL has been explored in a number of prior works. Kolter and Ng (2009); Geist and Scherrer (2011); Hoffman et al. (2011); Painter-Wakefield and Parr (2012) studied *on-policy* evaluation with ℓ_1-regularization for temporal-difference (TD) learning but none of them come with a theoretical analysis. Liu et al. (2012) studied off-policy evaluation by regularized TD learning but only provided algorithmic convergence guarantee without statistical error analysis.

Ghavamzadeh et al. (2011); Geist et al. (2012) proposed Lasso-TD with finite-sample statistical analysis for estimating the value function in *Markov reward process*. In particular, they derived in-sample *prediction error* bound $\mathcal{O}((s\log(d)/\psi n)^{1/2})$ under ψ-minimum eigenvalue condition on the empirical feature gram matrix. Although this bound also has no polynomial dependency on d, in-sample prediction error generally can not be translated to the estimation error of target policy in the OPE problem and their bound can not characterize the distribution mismatch between behavior policy and target policy. On the other hand, no minimax lower bound has been investigated so far.

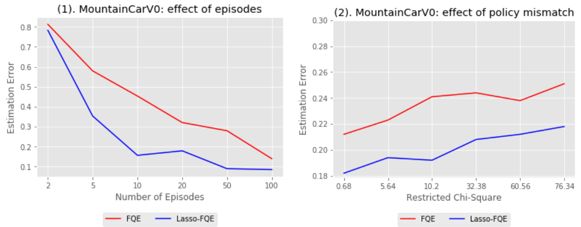
In addition, Farahmand et al. (2008; 2016) considered ℓ_2-regularization in fitted Q-iteration/policy iteration for policy optimization in a reproducing kernel Hilbert space, and finite-sample performance bounds for these algorithms were proved built on a coefficient concentration condition. Calandriello et al. (2014) developed Lasso fitted Q-iteration for sparse multi-task reinforcement learning and assumed a generative model (Kakade et al., 2003) for sampling transitions. Ibrahimi et al. (2012) derived a $\mathcal{O}(p\sqrt{T})$ regret bound in high-dimensional sparse linear quadratic systems where p is the dimension of the state space. More recently, Hao et al. (2020) studied the hardness of online reinforcement learning in fixed horizon, sparse linear Markov decision process.

Sparse linear regression. Sparse regression receives considerable attention in high-dimensional statistics in the past decade. Lasso (Tibshirani, 1996), is arguably the most widely used method to conduct sparse regression. Theoretical analysis of Lasso is well-studied in Zhao and Yu (2006); Bickel et al. (2009); Wainwright (2009). For a thorough review of Lasso as well as high-dimensional statistics, we refer the readers to Hastie et al. (2015); Wainwright (2019); Bühlmann and Van De Geer (2011). However, extending existing analysis from regression to batch RL is much more involved due to the complex optimization structure, non-i.i.d data collection, and covariate shift.

7. Experiment

We conducted experiments with a mountain car example. We use 800 radial basis functions for linear value function approximation. The number of episodes collected by behavior policies ranges from 2 to 100. We compare our Lasso-FQE with the standard FQE with ℓ_2 ridge regularization. For each algorithm we report the performance for the best regularization parameter λ in the range $\{0.02, 0.05, 0.1, 0.2, 0.5\}$. We pick a near-optimal policy as the target, and we measure the estimation error by $|\widehat{v} - v^*|/|v^*|$. All the results are averaged over 20 runs.

The left panel of Figure 7 shows that our Lasso-FQE clealy has smaller estimation error compared with FQE, proving the sparse feature selection is effective in a practical RL example. The right panel of Figure 7 demonstrates how the distribution mismatch (χ^2-divergence term) affects OPE error (with sample size fixed). We constructed multiple behavior policies with varying levels of ϵ-greedy noise, and plot their OPE error against their (restricted) χ^2-divergence from the target policy. The results confirm our theorems (Theorem 4.9&4.11) that the (restricted) chi-square divergence sharply determines the (sparse) OPE error.



8. Conclusion

In this work we focus on high-dimensional batch RL using sparse linear function approximation. While previous work in RL recognized the possibility of bringing tools from sparse learning to RL, they lacked a clean theoretical framework and formal results. By building on the strength of the linear DMDP framework, our result show that learning and planning in linear DMDPs can be done in the "feature space" even in the presence of sparsity and when only batch data is available.

Acknowledgements

Csaba Szepesvári gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii and NSERC. Mengdi Wang gratefully acknowledges funding from the U.S. National Science Foundation (NSF) grant CMMI1653435, Air Force Office of Scientific Research (AFOSR) grant FA9550-19-1-020, and C3.ai DTI.

References

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702, 2019.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020.

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted Q-iteration in continuous action-space MDPs. In *Advances in neural information processing systems*, pages 9–16, 2008.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

I. R. Bellman, R. Kalaba, and B. Kotkin. Polynomial approximation – a new computational technique in dynamic programming. *Math. Comp.*, 17(8):155–161, 1963.

Alexandre Belloni, Victor Chernozhukov, et al. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

Dimitri P. Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific, 1995.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

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

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

Daniele Calandriello, Alessandro Lazaric, and Marcello Restelli. Sparse multi-task reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 819–827, 2014.

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

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1): 129–159, 2001.

Yaqi Duan and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. *Internation Conference on Machine Learning*, 2020.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.

A.m. Farahmand, M. Ghavamzadeh, Cs. Szepesvári, and S. Mannor. Regularized policy iteration with nonparametric function spaces. *JMLR*, 17:1–66, 2016.

Amir massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized fitted Q-iteration: Application to planning. In *European Workshop on Reinforcement Learning*, pages 55–68. Springer, 2008.

Matthieu Geist and Bruno Scherrer. ℓ^1-penalized projected Bellman residual. In *European Workshop on Reinforcement Learning*, pages 89–101. Springer, 2011.

Matthieu Geist, Bruno Scherrer, Alessandro Lazaric, and Mohammad Ghavamzadeh. A Dantzig selector approach to temporal difference learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 347–354, 2012.

Mohammad Ghavamzadeh, Alessandro Lazaric, Rémi Munos, and Matthew Hoffman. Finite-sample analysis of Lasso-TD. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1177–1184, 2011.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1372–1383. JMLR. org, 2017.

Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. *arXiv preprint arXiv:2011.04018*, 2020.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Matthew W Hoffman, Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Regularized least squares temporal difference learning with nested ℓ^2 and ℓ^1 penalization. In *European Workshop on Reinforcement Learning*, pages 102–114. Springer, 2011.

Morteza Ibrahimi, Adel Javanmard, and Benjamin V Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems*, pages 2636–2644, 2012.

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

Sham Machandranath Kakade et al. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

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

J Zico Kolter and Andrew Y Ng. Regularization and feature selection in least-squares temporal difference learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 521–528, 2009.

Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of machine learning research*, 4(Dec):1107–1149, 2003.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

Nevena Lazic, Dong Yin, Mehrdad Farajtabar, Nir Levine, Dilan Gorur, Chris Harris, and Dale Schuurmans. A maximum-entropy approach to off-policy evaluation in average-reward mdps. *Conference on Neural Information Processing Systems*, 2020.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Lihong Li, Remi Munos, and Csaba Szepesvari. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616, 2015.

Bo Liu, Sridhar Mahadevan, and Ji Liu. Regularized off-policy TD-learning. In *Advances in Neural Information Processing Systems*, pages 836–844, 2012.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366, 2018.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May): 815–857, 2008.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, pages 2315–2325, 2019.

Christopher Painter-Wakefield and Ronald Parr. Greedy algorithms for sparse reinforcement learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 867–874, 2012.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *ICML'00 Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

Irina Rish and Genady Grabarnik. *Sparse Modeling: Theory, Algorithms, and Applications*. CRC Press, 2014.

Paul J Schweitzer and Abraham Seidmann. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110(2):568–582, 1985.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan and Claypool, 2010.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Masatoshi Uehara and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5): 2183–2202, 2009.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, pages 9665–9675, 2019.

Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004, 2019.

Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. *arXiv preprint arXiv:2001.10742*, 2020.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Ruiyi Zhang, Bo Dai, Lihong Li, and Dale Schuurmans. GenDICE: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020a.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. GradientDICE: Rethinking generalized offline estimation of stationary values. *arXiv preprint arXiv:2001.11113*, 2020b.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.