Statistica Sinica Preprint No: SS-2025-0457	
Title	Sparse Additive Off-policy Evaluation for Reinforcement Learning with Potentially Limited Number of Trajectories
Manuscript ID	SS-2025-0457
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0457
Complete List of Authors	Tuoyi Zhao, Chengchun Shi, Zhengling Qi and Lan Wang
Corresponding Authors	Lan Wang
E-mails	lxw611@miami.edu
Notice: Accepted author version.	

Statistica Sinica

Sparse Additive Off-Policy Evaluation for Reinforcement Learning with Potentially Limited Number of Trajectories

Tuoyi Zhao[1], Chengchun Shi[2], Zhengling Qi[3] and Lan Wang[4]

[1] *Yale University*

[2] *London School of Economics and Political Science*

[3] *George Washington University*

[4] *University of Miami*

Abstract: We develop a new framework for flexible, nonlinear, and interpretable off-policy evaluation for infinite-horizon reinforcement learning. To handle large state spaces and support transparent decision-making, we model the Q-function using a nonlinear function class with a sparse additive structure. We derive high-probability finite-sample error bounds for estimating the value function of a target policy and show that the bounds depend only logarithmically on the ambient dimension d, thereby alleviating the curse of dimensionality. In contrast to most existing theory for off-policy evaluation, which typically assumes access to many trajectories, our analysis guarantees accurate value estimation when either the number of trajectories or the time horizon is sufficiently large. In addition, we propose a group-sparsity-based feature screening procedure that identifies, with high probability, a reduced feature set containing all relevant covariates.

Numerical experiments demonstrate the effectiveness of the proposed approach.

Key words and phrases: High-dimensional data, Off-policy evaluation, Reinforcement learning, Sparsity.

1. INTRODUCTION

Off-policy evaluation (OPE) is a cornerstone of reinforcement learning (RL). Its goal is to estimate the performance of a target policy using offline data (historical interaction data) generated by a potentially different behavior policy (Sutton and Barto (2018)). Unlike online evaluation, where an agent directly interacts with the environment to test the policy, OPE avoids the risks and costs of deploying suboptimal or unsafe policies. This is especially important for domains where real-world interactions are expensive, risky, or ethically constrained. Examples include targeted advertising and marketing (Tang et al., 2013; Thomas et al., 2017), education (Mandel et al., 2014), and healthcare (Murphy et al., 2001; Murphy, 2003; Thomas et al., 2017), where poor decisions can lead to wasted resources, lost opportunities, or even harm to individuals.

Existing off-policy evaluation methods in RL can be broadly grouped into four categories: model-based approaches (Liu et al., 2018; Gottesman et al., 2019; Yin and Wang, 2020), importance sampling (IS)-based methods

(Liu et al., 2018; Wang et al., 2023), value function-based methods (Luckett et al., 2020; Liao et al., 2021), and doubly robust (DR) estimation methods (Jiang and Li, 2016; Thomas and Brunskill, 2016; Bibaut et al., 2021; Kallus and Uehara, 2022; Liao et al., 2022). Each class of methods strikes a different balance between bias and variance. However, the theoretical guarantees of most existing approaches rely on access to a sufficiently large number of trajectories, an assumption that is often unrealistic in real-world applications where data collection is costly or constrained. To address this challenge, Kallus and Uehara (2022) proposed a double reinforcement learning framework tailored to the limited-trajectory regime. While promising, their approach still requires estimating density ratios, which remains technically challenging and can be unstable in practice. More recently, Shi et al. (2022) investigated inference based on linear approximation of the Q-function and proved theoretical guarantees even with a limited number of trajectories, provided that the number of decision points is sufficiently large.

Offline RL with linear function approximation, but without exploiting sparsity structures, has been extensively studied (Bertsekas and Tsitsiklis, 1995; Lagoudakis and Parr, 2003; Munos, 2003; Alizadeh and Goldfarb, 2003; Tosatto et al., 2017; Duan et al., 2020; Jin et al., 2020; Shi et al.,

2022), among others. Building on this line of work, sparse learning in RL has also attracted significant attention. Early efforts focused on combining temporal-difference (TD) learning with ℓ_1 regularization in on-policy settings (Kolter and Ng, 2009; Geist and Scherrer, 2011; Hoffman et al., 2011; Painter-Wakefield and Parr, 2012), while Liu et al. (2012) extended these ideas to off-policy evaluation via regularized TD learning. However, these approaches largely lacked error bounds or consistency guarantees. To address this gap, Ghavamzadeh et al. (2011) proposed Lasso-TD and established in-sample guarantees for value function estimation, and Farahmand et al. (2016) developed a regularized policy iteration algorithm in nonparametric function spaces with accompanying statistical analysis. More recently, attention has shifted toward sparse linear MDPs: Hao et al. (021a) studied sparse linear approximation for offline RL and derived finite-sample error bounds that avoid polynomial dependence on the ambient dimension, thereby providing a theoretical foundation for the sample-efficiency benefits of sparsity. Their analysis was restricted to unstructured sparsity. Extensions include Hao et al. (021b), who analyzed the on-policy setting, and Golowich et al. (2024), who considered a broader class of ℓ_1-bounded linear MDPs and designed algorithms achieving near-optimal on-policy performance. Beyond Lasso-type approaches, Ma et al. (2026) proposed a

model-free variable selection framework using sequential knockoffs, which efficiently identifies minimal sufficient state representations.

In this work, we study infinite-horizon offline reinforcement learning, where the data-generating mechanism is modeled by a Markov decision process. Our focus is off-policy evaluation in large state spaces, and we extend semiparametric function approximation methods to incorporate group-wise sparsity structures. We are particularly interested in the high-dimensional regime where the state dimension d is very large and satisfies $d \gg \max\{n, T\}$, with n denoting the number of observed trajectories and T the number of observed time points.

We propose a new algorithm for estimating the cumulative discounted value of a target policy in the infinite-horizon setting that performs policy evaluation and feature selection simultaneously. Motivated by sparse additive modeling in supervised learning (Yuan and Lin, 2006; Ravikumar et al., 2009; Liu et al., 2010), our approach extends the infinite-horizon Bellman estimating framework of Shi et al. (2022), which is developed for low-dimensional state spaces. In our high-dimensional setting, the spline-expanded coefficient dimension can exceed nT. This necessitates group regularization and requires new theoretical analysis to establish uniform concentration bounds for the empirical Bellman scores under Markov de-

pendence, while simultaneously accounting for both the diverging spline dimension and the spline approximation error.

We derive high-probability finite-sample error bounds that depend on the ambient dimension d only through $\log d$ and d_0, the number of relevant features. In particular, the estimation error has no polynomial dependence on d, thereby mitigating the curse of dimensionality. Our theory guarantees accurate estimation when either the number of trajectories n or the number of decision points T is sufficiently large.

The proposed estimator remains consistent in probability as $T \to \infty$ even when n is fixed. Furthermore, we propose a group-sparsity-aware feature screening procedure that, with high probability, can identify a reduced set containing all relevant features. Our work thus significantly extends the tools and theory available for handling large state spaces for off-policy evaluation.

The rest of the paper is organized as follows. We introduce the notation and background in Section 2. Section 3 presents the sparsity-aware nonlinear sparse model and the estimation procedure. Section 4 derives the finite-sample error bounds for the value function under Markov errors. Numerical studies are reported in Section 5. The online supplement contains additional technical proofs.

2. PRELIMINARIES

2.1 Off-policy Evaluation for Reinforcement Learning

The dataset consists of n independent trajectories: $\{(X_{i,t}, A_{i,t}, R_{i,t}, X_{i,t+1}) : 1 \le i \le n,\ 1 \le t \le T-1\}$, where each trajectory $\{(X_{0,t}, A_{0,t}, R_{0,t}, X_{0,t+1}) : t \ge 0\}$ is generated according to a Markov decision process (MDP). At time t, a state $X_{0,t} \in \mathcal{X}^d$ is observed, an action $A_{0,t} \in \mathcal{A}$ is chosen according to the behavior policy $b(\cdot|X_{0,t})$, the system then transits to the next state $X_{0,t+1} \sim p(\cdot|X_{0,t}, A_{0,t})$ where $p(\cdot|\cdot,\cdot)$ denotes the transition probability (also called transition kernel or transition dynamics), and an immediate reward $R_{0,t}$ is received with conditional mean $r(x,a) = \mathbb{E}(R_{0,t} \mid X_{0,t} = x, A_{0,t} = a)$. It is standard to assume $|R_{i,t}| \le c_r$ for a positive constant c_r, $\forall\ i, t$.

We assume that the random trajectory satisfies the following Markov assumption (MA): for any $\mathcal{B} \subset \mathcal{X}^d$,

$$P\Big(X_{0,t+1} \in \mathcal{B} \,\Big|\, X_{0,t} = x, A_{0,t} = a, \{R_{0,j}\}_{0\le j<t}, \{X_{0,j}\}_{0\le j<t}, \{A_{0,j}\}_{0\le j<t}\Big)$$
$$= P\big(X_{0,t+1} \in \mathcal{B} \mid X_{0,t} = x, A_{0,t} = a\big),$$

and the conditional mean independence assumption (CMIA):

$$\mathbb{E}\Big(R_{0,t} \,|X_{0,t} = x, A_{0,t} = a, \{R_{0,j}\}_{0\le j<t}, \{X_{0,j}\}_{0\le j<t},$$
$$\{A_{0,j}\}_{0\le j<t}\Big) = \mathbb{E}\left(R_{0,t} \mid X_{0,t} = x, A_{0,t} = a\right).$$

If MA holds and $R_{0,t}$ is a deterministic function of $X_{0,t}$ and $A_{0,t}$, then CMIA is automatically satisfied (Shi et al., 2022).

Let π be a stationary policy of interest, which chooses the action $a \in \mathcal{A}$ with probability $\pi(a|x)$ given the state x. The performance of π is measured by the expected cumulative discounted reward, also called value function, defined as $V^\pi(x) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}^\pi \left(R_{0,t} \mid X_{0,0} = x\right)$, where $\gamma \in [0, 1)$ is the discount factor that balances the trade-off between the immediate and future rewards, $\mathbb{E}^\pi(\cdot \mid X_{0,0} = x)$ denotes the expectation with respect to the random trajectory when it begins with the initial state x and the actions along the trajectory are taken according to π. We also introduce the state-action value function or the Q-function:

$$Q^\pi(x, a) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}^\pi \left(R_{0,t} \mid X_{0,0} = x, A_{0,0} = a\right).$$

The Q function satisfies the Bellman consistency equations

$$Q^\pi(x, a) = r(x, a) + \gamma \mathbb{E}^\pi(Q^\pi(x', a') \mid x, a),$$

for all $(x, a) \in \mathcal{X}^d \times \mathcal{A}$. It is known that $V^\pi(x) = \sum_a \pi(a|x) Q^\pi(x, a)$. Our main goal is to estimate the cumulative discounted value of π using the offline trajectories data.

2.2 Q-function Modeling under Sparsity Constraint

In many applications, the number of features can be very large, potentially exceeding $\max\{n, T\}$. In such high-dimensional settings, existing methods may perform poorly: from a theoretical perspective, coverage conditions based on eigenvalues of feature matrices may fail because the corresponding matrices are no longer full rank. In this paper, we develop a new sparsity-aware Q-function modeling procedure to address the challenges associated with high-dimensional feature space.

Assumption 1 (**Sparse Q-function**)**.** Let π be the target policy. There exists an active feature set $\mathcal{K} \subseteq [d]$ with cardinality $|\mathcal{K}| \leq d_0 \ll d$. For each action $a \in \mathcal{A}$, there exists a subset $S_a \subseteq \mathcal{K}$ such that the Q-function depends only on the coordinates indexed by S_a, i.e.,

$$Q^\pi(x, a) = Q^\pi(x_{S_a}, a), \quad \forall\ (x, a) \in \mathcal{X} \times \mathcal{A}. \tag{2.1}$$

Because our goal is policy evaluation, we impose sparsity directly on the target-policy Q-function. Defined in terms of the original state coordinates, S_a identifies the characteristics that determine how long-run policy performance varies with the initial state and action a, while the estimated component functions describe these relationships. Variables outside S_a may affect rewards or transitions but provide no additional information about

the expected cumulative return conditional on the selected variables and initial action. This differs from the sparse linear MDP framework of Hao et al. (021a), which imposes sparsity on the transition structure relative to a prespecified feature dictionary. Our formulation instead captures sparsity in the Q-function after rewards and target-policy transition dynamics are combined.

The following proposition presents two alternative sufficient conditions for Assumption 1. These conditions serve only to illustrate structural settings that imply Assumption 1; the subsequent analysis imposes Assumption 1 directly and does not rely on either condition.

Proposition 1. *Assumption 1 holds if there exists an index set $\mathcal{K} \subseteq [d]$ with $|\mathcal{K}| \ll d$ such that either of the following conditions is satisfied:*

- ***(Sufficient Condition 1)*** *The reward function and the full transition kernel depend on the current state only through $x_{\mathcal{K}}$: $r(x, a) = r(x_{\mathcal{K}}, a)$ and $p(x' \mid x, a) = p(x' \mid x_{\mathcal{K}}, a)$.*

- ***(Sufficient Condition 2)*** *The reward function and the target policy depend only on $x_{\mathcal{K}}$: $r(x, a) = r(x_{\mathcal{K}}, a)$, $\pi(a \mid x) = \pi(a \mid x_{\mathcal{K}})$, and the transition of the active state satisfies $p(x'_{\mathcal{K}} \mid x, a) = p(x'_{\mathcal{K}} \mid x_{\mathcal{K}}, a)$.*

Under either condition, $Q^{\pi}(x, a) = Q^{\pi}(x_{\mathcal{K}}, a)$, so that $\mathcal{K}$ can serve as an

active feature set in Assumption 1.

Remark 1. Both sufficient conditions imply the joint-transition property

$$p^{\pi}(x'_{\mathcal{K}}, a' \mid x, a) = p^{\pi}(x'_{\mathcal{K}}, a' \mid x_{\mathcal{K}}, a), \tag{2.2}$$

where p^{π} denotes the conditional distribution of $(X'_{\mathcal{K}}, A')$ under the target policy. Together with $r(x, a) = r(x_{\mathcal{K}}, a)$, this property closes the Bellman recursion on the reduced state and implies $Q^{\pi}(x, a) = Q^{\pi}(x_{\mathcal{K}}, a)$.

Sufficient Condition 1 ensures (2.2) by requiring the full transition kernel of X' to depend on x only through $x_{\mathcal{K}}$, without restricting the target policy. Sufficient Condition 2 instead combines $p(x'_{\mathcal{K}} \mid x, a) = p(x'_{\mathcal{K}} \mid x_{\mathcal{K}}, a)$ and $\pi(a' \mid x') = \pi(a' \mid x'_{\mathcal{K}})$, which yields

$$p^{\pi}(x'_{\mathcal{K}}, a' \mid x, a) = p(x'_{\mathcal{K}} \mid x_{\mathcal{K}}, a)\pi(a' \mid x'_{\mathcal{K}}).$$

Thus, the conditions are not nested: the first imposes a stronger transition restriction, whereas the second adds a policy restriction to a reduced-state transition condition. This policy restriction is only sufficient for verifying (2.2); it is not used in the subsequent analysis, which relies directly on Assumption 1.

Property (2.2) identifies a low-dimensional, decision-relevant subsystem: under the target policy, the joint transition of $(X'_{\mathcal{K}}, A')$ depends on the current state only through $x_{\mathcal{K}}$. Thus, the Bellman continuation can be

expressed using the reduced state $X_{\mathcal{K}}$. The goal is to reduce the state information needed for policy evaluation, not to obtain a sparse representation of the full transition kernel. This perspective is related to the minimum sufficient MDP of Ma et al. (2026), although we do not require $\mathcal{K}$ to be minimal. In contrast, Hao et al. (021a) retain the full state space and impose sparsity on the transition representation relative to a prespecified feature map. Property (2.2) is only sufficient for verifying Assumption 1; the subsequent analysis imposes that assumption directly.

2.3 Notation

We next introduce some notations. Denote $[d] = \{1, \ldots, d\}$. For a d-dimensional vector $v = (v_1, \ldots, v_d)^{\mathrm{T}}$, let $\|v\|_1 = \sum_{i=1}^d |v_i|$, $\|v\|_2^2 = \sum_{i=1}^d |v_i|^2$ and $\|v\|_\infty = \max_i |v_i|$. Let $(v, w)_\oplus = (v_1, \ldots, v_d, w_1, \ldots, w_d)^{\mathrm{T}}$ denote the concatenation between vectors. For an arbitrary matrix A, we define the matrix norms: $\|A\|_{max} = \max_{i,j} |A_{i,j}|$ and $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$. For any real-valued bounded function $f(x)$ on $\mathcal{X}$, $||f(x)||_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. For any set $\mathcal{S}$, $\mathcal{S}^c$ denotes its complement, and $|\mathcal{S}|$ denotes its cardinality. In the case $\mathcal{S}$ is an index set, $v_{\mathcal{S}}$ denotes the $|\mathcal{S}|$-dimensional sub-vector $(v_i)_{i\in\mathcal{S}}$. For two generic sequences $\{a(n)\}_{n\geq 1}$ and $\{b(n)\}_{n\geq 1}$, $a(n) \lesssim b(n)$ means that there exists some positive constant c such that $a(n) \leq cb(n)$,

$\forall\ n$.

3. NONLINEAR SPARSE MODELING AND ESTIMATION FOR OFF-POLICY EVALUATION

3.1 Sparse Additive Modeling of the Q function

Let the state $x = (x_1, \ldots, x_d)^\top \in \mathcal{X}^d$. We assume a compact state space and, without loss of generality, take $\mathcal{X}^d = [0,1]^d$. We consider a high-dimensional setting where $d \gg \max\{n, T\}$ and focus on binary actions, but the method extends straightforwardly to any finite action space. To balance flexibility, interpretability, and statistical efficiency for high-dimensional RL, we adopt an additive modeling framework:

$$Q^\pi(x,a) = \alpha_a^* + \sum_{k=1}^{d} q_{k,a}^\pi(x_k), \quad a \in \{0,1\}, \tag{3.3}$$

where α_a^* are unknown constants and $q_{k,a}^\pi(x_k)$ are unknown univariate functions. This additive structure captures nonlinear effects of individual state components without requiring the estimation of a fully nonparametric multivariate function, thereby mitigating the curse of dimensionality.

Under the sparse Q-function assumption (Assumption 1),

$$Q^\pi(x,a) = Q^\pi(x_{S_a}, a) = \alpha_a^* + \sum_{s \in S_a} q_{s,a}^\pi(x_s), \quad a \in \{0,1\}, \tag{3.4}$$

where S_a is the active feature set associated with action a. For identifiability, we assume $\mathbb{E}_{x\sim\mathbb{G}}\{q^{\pi}_{s,a}(x_s)\} = 0$ for a reference distribution $\mathbb{G}$ on $\mathcal{X}^d$ (e.g., the state visitation distribution).

We approximate each $q^{\pi}_{s,a}(x_s)$ by a linear combination of basis functions and treat the basis coefficients associated with each state variable x_s as a group. The resulting representation is linear in the spline coefficients while retaining a flexible nonlinear dependence on the state variables, with the spline approximation error explicitly accounted for in our theory.

3.2 Assumptions for Function Approximation

Definition 1 (Smoothness)**.** For $\nu > 0$, let $p = \lceil \nu \rceil - 1$ and $\alpha = \nu - p \in (0, 1]$. A real-valued function h on $[0, 1]$ is said to be ν-smooth if it is p times continuously differentiable and

$$\sum_{j=0}^{p} \|h^{(j)}\|_{\infty} + \sup_{z\neq z'} \frac{|h^{(p)}(z') - h^{(p)}(z)|}{|z' - z|^{\alpha}} \leq c$$

for some constant $c > 0$. The collection of such functions is denoted by $\mathcal{H}_{\nu,c}$.

Assumption 2 (Function class)**.** There exist constants (ν, c) with $\nu \geq 3$ such that $q^{\pi}_{s,a}(\cdot) \in \mathcal{H}_{\nu,c}$ for all $s \in S_a$ and $a \in \{0, 1\}$.

Assumption 2 imposes smoothness conditions on $q^{\pi}_{s,a}(\cdot)$. We note that Assumption 2 can be relaxed to $\nu \geq 2$ for any $\epsilon > 0$ when d is small

and independent of n and T. We approximate each component function $q^{\pi}_{s,a}(x_s)$ using L centered B-spline contrast functions constructed from $L+1$ normalized B-spline basis functions of order $m = \lfloor \nu \rfloor$, denoted by $\mathbf{\Phi}_s(x_s)$. The detailed construction and its properties are provided in Appendix S2 of the online supplement. Since $\nu - 1 \leq m \leq \nu$, there exists a coefficient vector $\boldsymbol{\beta}^*_{s,a} \in \mathbb{R}^L$ such that for all $a \in \{0,1\}$ and $s \in S_a$, $\sup_{x_s \in \mathcal{X}} \left| q^{\pi}_{s,a}(x_s) - \Phi_s^{\top}(x_s)\boldsymbol{\beta}^*_{s,a} \right| \leq CL^{-m}$, for some constant $C > 0$ (e.g., Huang 1998, Schumaker 2007). Consequently,

$$\sup_{x \in \mathcal{X}} \left| Q^{\pi}(x,a) - \alpha^*_a - \sum_{s \in S_a} \Phi_s^{\top}(x_s)\boldsymbol{\beta}^*_{s,a} \right| \leq Cd_0 L^{-m}, \quad a \in \{0,1\}. \tag{3.5}$$

The number of basis functions L can diverge with n and/or T, alleviating the bias of sparse linear modeling while incorporating the group structure in estimation. We take $L \asymp (nT)^{1/(2m-1)}$ for theoretical derivations.

3.3 Regularized Estimation in High Dimensions

The Bellman equation for reinforcement learning (Sutton and Barto (2018)) implies that $\forall\ i, t$,

$$\mathbb{E}\left[R_{i,t} + \gamma \sum_{a \in \mathcal{A}} Q^{\pi}\left(X_{i,t+1}, a\right) \pi\left(a \mid X_{i,t+1}\right) - Q^{\pi}\left(X_{i,t}, A_{i,t}\right) \mid X_{i,t}, A_{i,t} \right] = 0. \tag{3.6}$$

Let $\boldsymbol{\beta}^*_{s,a} = 0$ for $s \notin S_a$. Define the basis functions $\Phi(x) = (1, \Phi_1(x_1), \ldots, \Phi_d(x_d))_{\oplus}$,

where $\oplus$ denotes concatenation of vectors. We approximate $Q^\pi(x,a)$ by

$$Q^*_\pi(x,a) = \Phi(x)^T(\alpha^*_a, \boldsymbol{\beta}^*_{1,a}, \ldots, \boldsymbol{\beta}^*_{d,a})_\oplus. \tag{3.7}$$

Then it follows from (3.6) that

$$\mathbb{E}\left[\left\{R_{i,t} + \gamma \sum_{a\in\mathcal{A}} Q^*_\pi\left(X_{i,t+1}, a\right)\pi\left(X_{i,t+1}, a\right) - Q^*_\pi\left(X_{i,t}, A_{i,t}\right)\right\} \mid X_{i,t}, A_{i,t}\right] \approx 0. \tag{3.8}$$

To introduce the proposed estimation procedure, we next introduce some new notation. Let $\boldsymbol{\alpha}^* = (\alpha^*_0, \alpha^*_1)^{\mathrm{T}}$. Consider the $(2Ld)$-dimensional vector $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^*_{1,0}, \ldots, \boldsymbol{\beta}^*_{d,0}, \boldsymbol{\beta}^*_{1,1}, \ldots, \boldsymbol{\beta}^*_{d,1})_\oplus$. Here and in the sequel, we drop π from the superscript in defining $\boldsymbol{\alpha}^*$, $\boldsymbol{\beta}^*$ and other related notation for simplicity. We introduce the group index G_s to denote the set of indices associated with x_s. This enables us to rewrite $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^*_{G_1}, \ldots, \boldsymbol{\beta}^*_{G_{2d}})_\oplus$, that is, $\boldsymbol{\beta}^*_{G_s} = \boldsymbol{\beta}^*_{s,0}$, $\boldsymbol{\beta}^*_{G_{s+d}} = \boldsymbol{\beta}^*_{s,1}$, for $1 \le s \le d$. Similarly, for $1 \le s \le d$, denote the vector $\boldsymbol{\xi}_{G_s}(x,a) = \Phi_s(x_s)\mathbb{I}(a=0)$, $\boldsymbol{\xi}_{G_{s+d}}(x,a) = \Phi_s(x_s)\mathbb{I}(a=1)$, $\boldsymbol{U}_{G_s}(x,a) = \Phi_s(x_s)\pi(0|x)$ and $\boldsymbol{U}_{G_{s+d}}(x,a) = \Phi_s(x_s)\pi(1|x)$, where $\mathbb{I}(\cdot)$ is the indicator function. Furthermore, we slightly abuse the notation and let $\Phi_0(\cdot) = 1$, $\boldsymbol{\xi}_{G_0}(x,a) = (\mathbb{I}(a=0), \mathbb{I}(a=1))^{\mathrm{T}}$, and let $\boldsymbol{U}_{G_0}(x,a) = (\pi(0|x), \pi(1|x))^{\mathrm{T}}$.

When estimating the Q^π function and evaluating the relevance of specific features, the group structure will be taken into account. Consider the short-hand notation $\boldsymbol{\xi}_{G_k,i,t} = \boldsymbol{\xi}_{G_k}(X_{i,t}, A_{i,t})$ and $\boldsymbol{U}_{G_k,i,t} = \boldsymbol{U}_{G_k}(X_{i,t})$,

$0 \leq k \leq 2d$. Define

$$\begin{aligned} \boldsymbol{\xi}_{i,t} &= \boldsymbol{\xi}(X_{i,t}, A_{i,t}) = (\boldsymbol{\xi}_{G_0,i,t}, \boldsymbol{\xi}_{G_1,i,t}, \ldots, \boldsymbol{\xi}_{G_{2d},i,t})_{\oplus}, \\ \boldsymbol{U}_{i,t} &= \boldsymbol{U}(X_{i,t}) = (\boldsymbol{U}_{G_0,i,t}, \boldsymbol{U}_{G_1,i,t}, \ldots, \boldsymbol{U}_{G_{2d},i,t})_{\oplus}. \end{aligned} \tag{3.9}$$

Note that $\boldsymbol{\xi}_{i,t}$ only depends on $X_{i,t}$ and $A_{i,t}$. It follows from (3.8) that

$$\mathbb{E}\left[\boldsymbol{\xi}_{i,t}\left\{R_{i,t} - (\boldsymbol{\xi}_{i,t} - \gamma \boldsymbol{U}_{i,t+1})^{\mathrm{T}}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)\right\}\right] \approx 0. \tag{3.10}$$

A similar set of estimating equations was recently considered in (Shi et al., 2022) for the case where d is small and fixed. In our setting $d \gg \max\{n, T\}$, the above set of estimating equations is ill-posed as the number of unknown parameters may be much larger than the number of equations.

To address the challenge of high-dimensional parameter estimation and to explicitly incorporate the group-wise sparsity structure, we introduce the following group Dantzig type procedure for estimating $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$. Specifically, the proposed estimator $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ are obtained by solving the following

constrained optimization problem:

$$\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}} = \arg\min_{\alpha,\boldsymbol{\beta}} \sum_{k=1}^{2d} \|\boldsymbol{\beta}_{G_k}\|_2, \quad \text{such that}$$
$$\left\| \sum_{i=1}^{n} \sum_{t=0}^{T-1} \boldsymbol{\xi}_{G_0,i,t} \left\{ R_{i,t} - (\boldsymbol{\xi}_{i,t} - \gamma \boldsymbol{U}_{i,t+1})^{\mathrm{T}} (\boldsymbol{\alpha}, \boldsymbol{\beta})_{\oplus} \right\} \right\|_2 \leqslant \frac{nT\lambda}{\sqrt{L}},$$
$$\text{and for } 1 \leq k \leq 2d, \tag{3.11}$$
$$\left\| \sum_{i=1}^{n} \sum_{t=0}^{T-1} \boldsymbol{\xi}_{G_k,i,t} \left\{ R_{i,t} - (\boldsymbol{\xi}_{i,t} - \gamma \boldsymbol{U}_{i,t+1})^{\mathrm{T}} (\boldsymbol{\alpha}, \boldsymbol{\beta})_{\oplus} \right\} \right\|_2 \leqslant nT\lambda.$$

Lemma 4 in the Appendix shows that the oracle coefficient vector $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)_{\oplus}$ satisfies the constraints with high probability. Specifically, when the empirical Bellman score is evaluated at $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)_{\oplus}$, it decomposes into a centered stochastic fluctuation and a spline-approximation term that admits a deterministic bound. After normalization by nT, the stochastic term is controlled uniformly over the coefficient groups at the rate $\lambda \asymp \sqrt{L}\log(nT \vee d)/\sqrt{nT}$, whereas the spline-approximation term is bounded by $Cd_0L^{1/2-m}$. To ensure that the latter is asymptotically negligible relative to λ, a convenient sufficient condition is $d_0\sqrt{nT}/L^m = o(1)$. Under the choice $L \asymp (nT)^{1/(2m-1)}$, this condition reduces to $d_0 = o\left((nT)^{1/(4m-2)}\right)$.

Given the estimators $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$, we estimate the Q-function or the value function at a given state x as follows. Let $\mathbb{G}$ denote the reference distribution of x, and let $\mathbb{G}_s$ denote the corresponding marginal distribution of the

feature x_s. Let $\widehat{q}_{s,a}(x_s) = \Phi_s^{\mathrm{T}}(x_s)\widehat{\boldsymbol{\beta}}_{s,a}$. and $\tilde{\alpha}_a = \widehat{\alpha}_a + \sum_{s=1}^{d} \int_{x \in \mathcal{X}^d} \widehat{q}^{\pi}_{s,a}(x) d\mathbb{G}$. Our estimator of $q^{\pi}_{s,a}(x_s)$ is given by

$$\widehat{q}^{\pi}_{s,a}(x_s) = \boldsymbol{\Phi}_s^{\top}(x_s)\widehat{\boldsymbol{\beta}}_{s,a}, \qquad \widetilde{\alpha}_a = \widehat{\alpha}_a + \sum_{s=1}^{d} \int_{\mathcal{X}} \widehat{q}^{\pi}_{s,a}(z)\, dG_s(z),$$

and

$$\widetilde{q}^{\pi}_{s,a}(x_s) = \widehat{q}^{\pi}_{s,a}(x_s) - \int_{\mathcal{X}} \widehat{q}^{\pi}_{s,a}(z)\, dG_s(z).$$

Correspondingly, we estimate $Q^*_{\pi}(x, a)$ by $\widehat{Q}^{\pi}(x, a) = \tilde{\alpha}_a + \sum_{s=1}^{d} \tilde{q}^{\pi}_{s,a}(x_s)$, which can also be expressed as

$$\widehat{Q}^{\pi}(x, a) = \Phi^{\mathrm{T}}(x)(\widehat{\alpha}_a, \widehat{\boldsymbol{\beta}}_{1,a}, \ldots, \widehat{\boldsymbol{\beta}}_{d,a})_{\oplus}, \quad a \in \{0, 1\}.$$

Remark 2. Although (3.11) draws on high-dimensional additive regression (Yuan and Lin, 2006; Candes and Tao, 2007; Liu et al., 2010), its analysis must accommodate Markov-dependent Bellman scores rather than independent regression errors. The estimator is also partially penalized: $(\alpha_0, \alpha_1)^{\top}$ are intrinsic model parameters and remain unpenalized, while the spline coefficient groups are regularized. Unlike a conventional regression intercept, these parameters are retained explicitly rather than removed by centering. This formulation also permits prespecified important covariates to remain unpenalized.

Remark 3. The formulation extends naturally to other group structures in RL. Groups may reflect scientifically meaningful feature sets, such as genes sharing biological functions or pathways; collections of indicators encoding categorical variables; or learned feature representations from neural networks. The spline coefficients associated with each state variable are one instance of this general structure. Treating each group as a unit can improve value estimation and interpretability by identifying the feature groups most relevant to policy evaluation.

Remark 4. We observe that the objective function in (3.11) is equivalent to

$$\min_{\boldsymbol{\tau},\boldsymbol{\alpha},\boldsymbol{\beta}} \mathbf{1}^\top \boldsymbol{\tau} \qquad \text{subject to} \qquad \|\boldsymbol{\beta}_{G_k}\|_2 \le \tau_k, \quad 1 \le k \le 2d,$$

where $\mathbf{1} = (1, 1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^{2d}$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{2d})^{\mathrm{T}} \in \mathbb{R}^{2d}$. With the same constraints as in (3.11), this becomes a second-order cone programming (SOCP) problem (Lobo et al., 1998; Alizadeh and Goldfarb, 2003).

4. FINITE-SAMPLE ERROR BOUNDS

4.1 Additional Technical Conditions

The Markov chain $\{X_{0,t}\}_{t\ge 0}$ is assumed to have a unique invariant distribution with some density function $\mu(\cdot)$. The density function μ and the initial

density (i.e., density of the initial state) μ_0 are both uniformly bounded away from 0 and ∞. This is a standard assumption in the RL literature for studying off-policy estimation.

We next introduce the restricted minimum eigenvalue condition. Recall that S_0 and S_1 denote the set of active features corresponding to $a \in \{0, 1\}$, respectively. Let $\mathcal{K} = S_0 \cup S_1$ and denote its cardinality $|\mathcal{K}| = d_0$. We refer to d_0, the total number of active features, as the sparsity size. Let $\mathcal{J}$ denote the index set corresponding to all active features, and let $\mathcal{J}^C$ denote its complement. Define the set

$$\mathbb{C} = \left\{ \delta \in \mathbb{R}^{2(dL+1)} : \delta \neq \mathbf{0}, \sum_{G_k \subset \mathcal{J}^C, k \neq 0} \|\delta_{G_k}\|_2 \leq \sum_{G_k \subset \mathcal{J}} \|\delta_{G_k}\|_2 \right\}.$$

It can be shown that $(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\oplus}$ belongs to the set $\mathbb{C}$ with high probability, see the online supplement for the proof.

Definition 2 (Restricted minimum eigenvalue)**.** The restricted minimum eigenvalue of a matrix K with respect to the set $\mathbb{C}$ is defined as $\lambda^r_{min}(K, \mathbb{C}) = \min_{\delta \in \mathbb{C}} \frac{\delta^{\mathrm{T}} K \delta}{\|\delta\|_2^2}$.

Let $\Lambda = \sum_{t=0}^{T-1} \mathbb{E}\left\{ \boldsymbol{\xi}_{0,t} \boldsymbol{\xi}_{0,t}^{\top} - \gamma^2 \boldsymbol{u}\left(X_{0,t}, A_{0,t}\right) \boldsymbol{u}^{\top}\left(X_{0,t}, A_{0,t}\right) \right\}$, where $\boldsymbol{u}(x, a) = \mathbb{E}\left\{ \boldsymbol{U}\left(X_{0,1}\right) \mid X_{0,0} = x, A_{0,0} = a \right\}$. The restricted eigenvalue condition is stated as follows:

Assumption 3. There exists some $\bar{c} > 0$ such that $\lambda^r_{min}(\Lambda, \mathbb{C}) \geq \bar{c} T$.

Remark 5 (Coverage interpretation of Assumption 3)**.** Assumption 3 is a restricted eigenvalue condition analogous to those used in high-dimensional regression (e.g., Bickel et al., 2009). It can also be interpreted as a feature-level coverage condition for the behavior policy. In linear off-policy evaluation, coverage is typically characterized through the covariance matrix of the features observed under the behavior policy (Duan et al., 2020; Hao et al., 021a). Our condition plays a similar role but additionally accounts for the target-policy-weighted next-state features appearing in the Bellman equation. Specifically, $\boldsymbol{U}_{0,t+1}$ averages the spline features at the next state $X_{0,t+1}$ over the possible next actions according to the target policy, while $\boldsymbol{u}(X_{0,t}, A_{0,t})$ is its conditional expectation given the current state and action.

Let μ_t denote the marginal density of $X_{0,t}$ under the behavior policy and define $\bar{\mu}_T(x) = T^{-1}\sum_{t=0}^{T-1}\mu_t(x)$. Then

$$\frac{\Lambda}{T} = \int_{\mathcal{X}^d}\sum_{a\in\mathcal{A}}\left\{\boldsymbol{\xi}(x,a)\boldsymbol{\xi}^\top(x,a) - \gamma^2\boldsymbol{u}(x,a)\boldsymbol{u}^\top(x,a)\right\} b(a\mid x)\bar{\mu}_T(x)\,dx.$$

Thus, Assumption 3 is equivalent to requiring, for every $\delta \in \mathbb{C}$,

$$\mathbb{E}_{\bar{\mu}_T,b}\left[\{\delta^\top\boldsymbol{\xi}(X,A)\}^2 - \gamma^2\{\delta^\top\boldsymbol{u}(X,A)\}^2\right] \geq \bar{c}\|\delta\|_2^2.$$

The behavior policy must therefore generate sufficient variation in every relevant sparse spline direction after accounting for the discounted target-

policy continuation features. The condition may fail if the behavior policy rarely selects a relevant action or visits regions where a relevant basis function varies, or if the current and continuation features are nearly collinear along a sparse direction. No separate pointwise overlap or target-policy eigenvalue condition is imposed; the required coverage is encoded through Λ in the feature directions entering the Bellman equation.

The factor T in Assumption 3 reflects that Λ sums the population matrices over T time points: $\lambda^r_{\min}\left(\frac{\Lambda}{T}, \mathbb{C}\right) \geq \bar{c}$. The normalized matrix Λ/T may nevertheless vary with T through $\bar{\mu}_T$. If $\bar{\mu}_T$ converges and the limiting Bellman matrix has a restricted eigenvalue bounded away from zero, the assumption holds for all sufficiently large T. However, increasing T cannot remedy a lack of coverage under the long-run behavior distribution.

Unlike fixed-dimensional analyses, which typically impose an ordinary minimum eigenvalue condition, we require the bound only over the sparse cone $\mathbb{C}$, since the spline-expanded dimension may exceed nT.

Our theory in Section 4.2 implies that the value function of interest can be consistently estimated when either n or T is large. Assumption 4 is used only to control the discrepancy between the empirical and population covariance matrices uniformly over their entries when T is large; it is not needed when n is large.

Assumption 4. The Markov chain $\{X_{0,t}\}_{t\geq 0}$ is exponentially β mixing.

Remark 6. When $X_{0,t}$ is stationary, Assumption 4 can be reduced to the assumption that the Markov chain $X_{0,t}$ is geometrically ergodic (Bradley, 2005). Assumption 4 does not imply the population coverage condition; rather, it controls the concentration of empirical Bellman matrices around their population counterparts. Under exponential β-mixing, $\beta(k) \leq C_\beta \rho^k$ for some $\rho \in (0,1)$, the proof uses blocks of length of order $\log(nT \vee d)/|\log \rho|$ to make the coupling error negligible. Thus, observations from a long trajectory contribute information, although temporal dependence reduces the effective number of approximately independent observations.

4.2 Finite-sample Error Bound for the Estimated Coefficients

Theorem 5 below provides high-probability finite error bounds for the proposed estimator $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})_\oplus$. Let $\widehat{\boldsymbol{\kappa}} = (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*, \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_\oplus$.

Theorem 5. *Under Assumption 1-4, we have*

$$||\widehat{\boldsymbol{\kappa}}||_1 \lesssim \frac{\log(nT \vee d)}{\sqrt{nT}} * d_0 L, \tag{4.12}$$

and

$$||\widehat{\boldsymbol{\kappa}}||_2 \lesssim \frac{\log(nT \vee d)}{\sqrt{nT}} * \sqrt{d_0 L}, \tag{4.13}$$

with probability at least $1 - \exp\{-c \log(nT \vee d)\}$ *for some constant* $c > 0$.

The ambient dimension d enters the finite-sample error bounds only through the logarithmic factor $\log(nT \vee d)$. To illustrate this dependence, suppose $d \asymp (nT)^{\kappa}$ for some fixed $\kappa > 0$. Under $L \asymp (nT)^{1/(2m-1)}$, $\|\widehat{\boldsymbol{\kappa}}\|_1 \lesssim d_0 \log(nT)(nT)^{-(2m-3)/(4m-2)}$. Consequently, if $d_0 = o\{(nT)^{1/(4m-2)}\}$, then $\|\widehat{\boldsymbol{\kappa}}\|_1 = o\left\{\log(nT)(nT)^{-(m-2)/(2m-1)}\right\} = o(1)$ for $m \geq 3$. Similarly, $\|\widehat{\boldsymbol{\kappa}}\|_2 \lesssim \sqrt{d_0}\log(nT)(nT)^{-(m-1)/(2m-1)}$, and ℓ_2 consistency requires only $\sqrt{d_0}\log(nT) = o\left\{(nT)^{(m-1)/(2m-1)}\right\}$. The estimator thus remains consistent when the ambient dimension grows at any fixed polynomial rate, provided that the sparsity size satisfies the stated growth conditions. The polynomial-growth regime is used only to illustrate the dimensional dependence; the theorem applies to any growth rate of d satisfying its assumptions and yielding vanishing upper bounds. When d and d_0 are fixed, a spectral-norm argument replaces the high-dimensional max-norm analysis and relaxes Assumption 2 to $\nu \geq 2$, or equivalently, $m = \lfloor \nu \rfloor \geq 2$.

4.3 Finite-sample Error Bound of Value Function

Given the target policy π, we estimate the expected cumulative discounted reward $V^{\pi}(x)$ by $\widehat{V}^{\pi}(x) = \mathbb{E}^{\pi}\{\Phi(x)^T(\widehat{\alpha}_a, \widehat{\boldsymbol{\beta}}_{1,a}, \ldots, \widehat{\boldsymbol{\beta}}_{d,a})_{\oplus}\}$. The next theorem provides a finite-sample error bound for $\widehat{V}^{\pi}(x)$.

Theorem 6. *Assume the conditions in Theorem 5 are satisfied. With prob-*

ability at least $1-\exp\{-c\log(nT\vee d)\}$*, we have*

$$\sup_{x\in\mathcal{X}^d}|\widehat{V}^\pi(x)-V^\pi(x)|\lesssim d_0L\frac{\log(nT\vee d)}{\sqrt{nT}}. \tag{4.14}$$

If the density of distribution $\mathbb{G}$ *is uniformly bounded above for a deterministic policy* π*, we have*

$$|\mathbb{E}_{x\sim\mathbb{G}}\widehat{V}^\pi(x)-\mathbb{E}_{x\sim\mathbb{G}}V^\pi(x)|\lesssim d_0\sqrt{L}\frac{\log(nT\vee d)}{\sqrt{nT}}. \tag{4.15}$$

The proof of the theorem relies on the following decomposition

$$\left|\widehat{V}^\pi(x)-V^\pi(x)\right|\le\left|\widehat{V}^\pi(x)-V^*_\pi(x)\right|+\left|V^*_\pi(x)-V^\pi(x)\right|,$$

where $V^*_\pi(x)=\sum_{a\in\mathcal{A}}\left\{\alpha^*_a+\sum_{k=1}^d\Phi_k^\top(x_k)\boldsymbol{\beta}^*_{k,a}\right\}\pi(a|x)$, and the two terms on the right-hand side represent the estimation error and the spline approximation error, respectively. Standard B-spline approximation results imply that $\sup_{x\in\mathcal{X}^d}\left|V^\pi(x)-V^*_\pi(x)\right|\lesssim d_0L^{-m}$. Theorem 5 is then used to control the estimation error. In particular, it yields the uniform bound $\sup_{x\in\mathcal{X}^d}\left|\widehat{V}^\pi(x)-V^*_\pi(x)\right|\lesssim d_0L\frac{\log(nT\vee d)}{\sqrt{nT}}$. For the integrated value error, when $\mathbb{G}$ has a uniformly bounded density and π is deterministic, the integral properties of the B-spline basis yield the sharper bound $\left|\mathbb{E}_{x\sim\mathbb{G}}\widehat{V}^\pi(x)-\mathbb{E}_{x\sim\mathbb{G}}V^*_\pi(x)\right|\lesssim d_0\sqrt{L}\log(nT\vee d)/\sqrt{nT}$. Therefore,

$$\left|\mathbb{E}_{x\sim\mathbb{G}}\widehat{V}^\pi(x)-\mathbb{E}_{x\sim\mathbb{G}}V^\pi(x)\right|\lesssim d_0\sqrt{L}\frac{\log(nT\vee d)}{\sqrt{nT}}+d_0L^{-m}. \tag{4.16}$$

For $L \asymp (nT)^{1/(2m-1)}$, the approximation error is asymptotically negligible relative to the integrated estimation error since $L^{-m}\sqrt{nT}\{\sqrt{L}\log(nT \vee d)\}^{-1} = o(1)$. The same conclusion holds for the uniform error bound, whose estimation component is larger by a factor of $\sqrt{L}$. Thus, the spline approximation error can be absorbed into the respective estimation errors, yielding the simplified bounds stated in Theorem 6.

Remark 7. Our result differs from the high-probability guarantee of (Hao et al., 021a, Theorem 4.8) in how the trajectory length affects the confidence level. In their notation, the data consist of $N = KL$ observations from K independent trajectories, each of length L. Their sample-size condition requires

$$N \gtrsim \frac{Ls^2 \log(d^2/\delta)}{C_{\min}^2(\Sigma, s)} + \frac{\gamma^2 Ls \log(s/\delta)}{(1-\gamma)^2}.$$

Because $N = KL$, this condition is equivalent to a lower bound on the number of independent trajectories,

$$K \gtrsim \frac{s^2 \log(d^2/\delta)}{C_{\min}^2(\Sigma, s)} + \frac{\gamma^2 s \log(s/\delta)}{(1-\gamma)^2}.$$

Consequently, under their stated sufficient conditions, increasing the trajectory length L while holding K fixed does not permit the failure probability δ to converge to zero, even though the error bound on the corresponding good event decreases with the total sample size. In contrast, our result estab-

lishes long-trajectory consistency both in the magnitude of the estimation error and in the confidence level of the guarantee.

4.4 Feature Screening

Next, we propose a screening procedure to assess feature relevance. The procedure explicitly incorporates the group structure: if a feature is irrelevant, the entire group of coefficients associated with its basis functions is expected to be screened out, leading to more interpretable results. We define the set of selected features as

$$\widehat{\mathcal{K}} = \{1 \le s \le d : \|\widehat{\boldsymbol{\beta}}_{G_s}\|_2 > h \quad \text{or} \quad \|\widehat{\boldsymbol{\beta}}_{G_{s+d}}\|_2 > h\}, \tag{4.17}$$

where h is a predefined threshold. The effectiveness of this screening step relies on the marginal signal of active components not being dominated by estimation error, which vanishes under suitable conditions implied by our earlier error bounds. The following theorem provides a theoretical guarantee for the proposed screening procedure.

Theorem 7. *Suppose the conditions in Theorem 5 are satisfied and d_0 is fixed. If there exists a constant v such that $\max_{a\in\mathcal{A}} \mathbb{E}_{x\sim\mathbb{G}}(q^{\pi 2}_{s,a}(x)) \ge v$ for all $s \in \mathcal{K}$, then setting $h = \sqrt{v/(8c_m^2 L)}$, we have $P(\mathcal{K} \subseteq \widehat{\mathcal{K}}) \ge 1 - \exp\{-c\log(nT \vee d)\}$.*

5. NUMERICAL EXAMPLES

We compare our proposed estimator (denoted by GD, due to the group Dantzig structure used in estimation) with the SAVE estimator from Shi et al. (2022) and Lasso-FQE (denoted by LFQ) from Hao et al. (021a). All results are summarized in Appendix S5.

5.1 Experiment 1

In this model, we generate data from a Markovian decision process described below. The system evolves according to the following dynamics:

$$x_{t+1,1:d} = Ax_{t,1:d} + Ba_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2_{\text{trans}} I),$$
$$r_t = w^{\mathrm{T}} x_{t,1:4} + \rho\big(x_{t,1}x_{t,3} + x_{t,2}x_{t,4}\big) - ca_t + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \sigma^2_{\text{reward}}).$$

The transition matrix A is block diagonal, consisting of $d/2$ blocks of size 2×2, so that every pair of adjacent features is coupled. To generate each block, we first sample its four entries independently from Unif$[0.80, 0.90]$ and then rescale the entire block to have spectral radius 0.95. Consequently, $\rho(A) = 0.95$, ensuring the stability of the state process. The entries of the d-dimensional vector B are independently sampled from Unif$[-0.2, 0.2]$. For each value of d, A and B are generated once and held fixed across all replications. We initialize each trajectory at $\boldsymbol{x}_0 = \mathbf{0}$ and set $\sigma_{\text{trans}} = \sigma_{\text{reward}} = 0.1$.The d-dimensional vector B satisfies $\|B\|_\infty \leq 0.2$ and ensures that action can affect each state coordinate. The reward function is linear

in the first four features $x_{1:4}$ and the action a with coefficients $(w, c) = (0.9, -0.7, 0.9, 0.6, 0.2)$, $\rho = 0.7$ and $\sigma_{\text{reward}} = 0.1$. The behavior policy belongs to the logistic policy class and is defined over the first 4 features with coefficients $(1.2, -0.9, 0.6, 0.7)$. The target policy follows same class with coefficients $(1, -0.5, 1.5, -1)$.

We consider a small state with $d = 10$ and a larger state $d = 100 \gg \max\{n, T\}$, and vary the number of independent trajectories and the trajectory length over $n \in \{20, 30, 40, 50\}$ and $T \in \{20, 30, 40\}$. We set $\lambda = 0.001$, $L = 8$ and $\gamma = 0.8$, and report the results based on 100 independent runs. In each run, we evaluate the bias of $\widehat{V}(x)$ at $x = 0$, that is, $\left[\widehat{V}(x) - V_{\text{MC}}(x)\right]\big|_{x=0}$, where $V_{\text{MC}}(x)$ denotes the Monte Carlo estimate of the cumulative reward over a horizon of 100, averaged over 1000 repetitions. Tables 1 and 2 summarize the results for $d = 10$ and 100, respectively. For $d = 10$, GD remains nearly unbiased across all combinations of (n, T), with its standard deviationgenerally decreasing as either n or T increases. Its performance is comparable to that of the other methods in this setting. For $d = 100$, GD continues to produce small bias and low variability over the entire (n, T) grid. Its standard deviation generally improves as the available sample size increases. Lasso-FQE remains numerically stable, but has a substantial positive bias for shorter trajectories. However, SAVE is

substantially less stable in the high-dimensional setting. It occasionally produces extreme estimates, as reflected by the large standard deviations at $(n, T) = (30, 40)$, $(40, 30)$, and $(50, 30)$.

Overall, GD provides the most reliable performance across both the low-dimensional and high-dimensional cases.

5.2 Experiment 2: Cartpole

We consider the CartPole environment from OpenAI Gym and manually incorporated null variables into the state, with $d_0 = 4$ and $d = 80$. The null variables were generated from a standard normal distribution independently. To encourage the pole to remain balanced for as long as possible, we set the discount rate $\gamma = 0.98$. The target policy selects $A = 1$ with probability 0.5. The behavior policy belongs to the logistic policy class and is designed to favor pushing the cart to the right when the pole is leaning to the right: $b(1 \mid x) = 1/\{1 + \exp[-(x_3 + x_4)]\}$.

We evaluate $V(x)$ at $x = 0$. The true value function $V^\pi(x)|_{x=0} \approx 33.50$ is computed using Monte Carlo approximations with 10^5 independent trajectories. Each trajectory consists of $T = 500$ decision points. We set $\lambda = 0.001$ and $L = 8$.

The results are given in the online supplement to save space. We ob-

serve that GD performs satisfactorily across the considered combinations of (n, T). Lasso-FQE exhibits a systematic negative bias when T is small, but this bias decreases substantially as the trajectory length increases. SAVE is still substantially less stable. Although it performs well for some settings, its standard deviation can be very large, with particularly extreme behavior at $(n, T) = (80, 50)$.

Overall, GD delivers the most stable performance across the reported configurations, even though the true Q-function does not satisfy a purely additive structure.

Supplementary Material

Our supplementary material contains the technical derivations and detailed numerical results. It consists of four parts. The first part provides additional details on the sparsity-aware Q-functions and the properties of B-spline functions used in our analysis. The second part contains technical lemmas. Lemma 3 and 5 provide spectral results related to $\widehat{\boldsymbol{\Sigma}}$. Lemma 4 shows that the optimal model coefficients $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)_{\oplus}$ are feasible. The third part provides the proofs of our main results, Theorem 5, 6, and 7. The final part provides the numerical results corresponding to Section 5, which are omitted from the main text due to the page limit.

Acknowledgements

The research of Zhao was supported in part by R01HL158075, P50HD052120, and the Lambert Family Fellowship. The research of Wang was partially supported by NSF DMS-2610563.

References

Alizadeh, F. and D. Goldfarb (2003). Second-order cone programming. *Mathematical Programming 95*(1), 3–51.

Bertsekas, D. P. and J. N. Tsitsiklis (1995). Neuro-dynamic programming: an overview. In *Proceedings of 1995 34th IEEE conference on decision and control*, Volume 1, pp. 560–564. IEEE.

Bibaut, A., M. Petersen, N. Vlassis, M. Dimakopoulou, and M. van der Laan (2021). Sequential causal inference in a single world of connected units.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705 – 1732.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys 2*, 107–144.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313 – 2351.

REFERENCES

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

Farahmand, A.-m., M. Ghavamzadeh, C. Szepesvári, and S. Mannor (2016). Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research 17*(1), 4809–4874.

Geist, M. and B. Scherrer (2011). l1-penalized projected Bellman residual. In *European Workshop on Reinforcement Learning*, pp. 89–101. Springer.

Ghavamzadeh, M., A. Lazaric, R. Munos, and M. Hoffman (2011). Finite-sample analysis of Lasso-TD. In *International Conference on Machine Learning*.

Golowich, N., A. Moitra, and D. Rohatgi (2024). Exploring and learning in sparse linear mdps without computationally intractable oracles. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 183–193.

Gottesman, O., Y. Liu, S. Sussex, E. Brunskill, and F. Doshi-Velez (2019). Combining parametric and nonparametric models for off-policy evaluation. In *International Conference on Machine Learning*, pp. 2366–2375. PMLR.

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

Hao, B., T. Lattimore, C. Szepesvári, and M. Wang (2021b). Online sparse reinforcement

learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 316–324. PMLR.

Hoffman, M. W., A. Lazaric, M. Ghavamzadeh, and R. Munos (2011). Regularized least squares temporal difference learning with nested l2 and l1 penalization. In *European Workshop on Reinforcement Learning*, pp. 102–114. Springer.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics 26*(1), 242–272.

Jiang, N. and L. Li (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR.

Jin, C., Z. Yang, Z. Wang, and M. I. Jordan (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR.

Kallus, N. and M. Uehara (2022). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Operations Research 70*(6), 3282–3302.

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

Lagoudakis, M. G. and R. Parr (2003). Least-squares policy iteration. *The Journal of Machine Learning Research 4*, 1107–1149.

Liao, P., P. Klasnja, and S. Murphy (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. *Journal of the American Statistical Association 116*(533), 382–391.

Liao, P., Z. Qi, R. Wan, P. Klasnja, and S. A. Murphy (2022). Batch policy learning in average reward markov decision processes. *Annals of statistics 50*(6), 3364–3387.

Liu, B., S. Mahadevan, and J. Liu (2012). Regularized off-policy TD-learning. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, Volume 25. Curran Associates, Inc.

Liu, H., J. Zhang, X. Jiang, and J. Liu (2010, 13–15 May). The group Dantzig selector. In Y. W. Teh and M. Titterington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Volume 9 of *Proceedings of Machine Learning Research*, Chia Laguna Resort, Sardinia, Italy, pp. 461–468. PMLR.

Liu, Q., L. Li, Z. Tang, and D. Zhou (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 31. Curran Associates, Inc.

Liu, Y., O. Gottesman, A. Raghu, M. Komorowski, A. A. Faisal, F. Doshi-Velez, and E. Brunskill (2018). Representation balancing MDPs for off-policy policy evaluation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 31. Curran Associates, Inc.

Lobo, M. S., L. Vandenberghe, S. Boyd, and H. Lebret (1998). Applications of second-order cone programming. *Linear Algebra and its Applications 284*(1-3), 193–228.

Luckett, D. J., E. B. Laber, A. R. Kahkoska, D. M. Maahs, E. Mayer-Davis, and M. R. Kosorok (2020). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association 115*(530), pp. 692–706.

Ma, T., J. Zhu, H. Cai, Z. Qi, Y. Chen, C. Shi, and E. B. Laber (2026). Sequential knockoffs for variable selection in reinforcement learning. *Journal of the American Statistical Association 0*(ja), 1–29.

Mandel, T., Y.-E. Liu, S. Levine, E. Brunskill, and Z. Popovic (2014). Offline policy evaluation across representations with applications to educational games. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, pp. 1077–1084.

Munos, R. (2003). Error bounds for approximate policy iteration. In *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 560–567.

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

Murphy, S. A., M. J. van der Laan, J. M. Robins, and C. P. P. R. Group (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association 96*(456), 1410–1423.

Painter-Wakefield, C. and R. Parr (2012). Greedy algorithms for sparse reinforcement learning. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 867–874.

REFERENCES

Ravikumar, P., J. Lafferty, H. Liu, and L. Wasserman (2009). Sparse additive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 71*(5), 1009–1030.

Schumaker, L. (2007). *Spline functions: basic theory*. Cambridge University Press.

Shi, C., S. Zhang, W. Lu, and R. Song (2022). Statistical inference of the value function for reinforcement learning in infinite-horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(3), 765–793.

Sutton, R. S. and A. G. Barto (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book.

Tang, L., R. Rosales, A. Singh, and D. Agarwal (2013). Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 1587–1594.

Thomas, P. and E. Brunskill (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR.

Thomas, P., G. Theocharous, M. Ghavamzadeh, I. Durugkar, and E. Brunskill (2017). Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 31, pp. 4740–4745.

Tosatto, S., M. Pirotta, C. d'Eramo, and M. Restelli (2017). Boosted fitted Q-iteration. In *International Conference on Machine Learning*, pp. 3434–3443. PMLR.

REFERENCES

Wang, J., Z. Qi, and R. K. W. Wong (2023). Projected state-action balancing weights for offline reinforcement learning. *The Annals of Statistics 51*(4), 1639 – 1665.

Yin, M. and Y.-X. Wang (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958. PMLR.

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

Yale University

E-mail: tuoyi.zhao@yale.edu

London School of Economics and Political Science

E-mail: c.shi7@lse.ac.uk

George Washington University

E-mail: qizhengling@email.gwu.edu

University of Miami

E-mail: lanwang@mbs.miami.edu