Sparse Offline Reinforcement Learning with Corruption Robustness

Nam Phuong Tran
University of Warwick

Andi Nika
MPI-SWS

Goran Radanovic
MPI-SWS

Long Tran-Thanh
University of Warwick

Debmalya Mandal
University of Warwick

Abstract

We investigate robustness to strong data corruption in offline sparse reinforcement learning (RL). In our setting, an adversary may arbitrarily perturb a fraction of the collected trajectories from a high-dimensional but *sparse* Markov decision process, and our goal is to estimate a near optimal policy. The main challenge is that, in the high-dimensional regime where the number of samples N is smaller than the feature dimension d, exploiting sparsity is essential for obtaining non-vacuous guarantees but has not been systematically studied in offline RL. We analyse the problem under *uniform coverage* and *sparse single-concentrability* assumptions. While Least Square Value Iteration (LSVI), a standard approach for robust offline RL, performs well under uniform coverage, we show that integrating sparsity into LSVI is unnatural, and its analysis may break down due to overly pessimistic bonuses. To overcome this, we propose actor–critic methods with sparse robust estimator oracles, which avoid the use of pointwise pessimistic bonuses and provide the first non-vacuous guarantees for sparse offline RL under single-policy concentrability coverage. Moreover, we extend our results to the contaminated setting and show that our algorithm remains robust under strong contamination. Our results provide the first non-vacuous guarantees in high-dimensional sparse MDPs with single-policy concentrability coverage and corruption, showing that learning near-optimal policy remains possible in regimes where traditional robust offline RL techniques may fail.

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

1 INTRODUCTION

Offline reinforcement learning (RL) aims to learn effective decision-making policies solely from previously collected data, without further interaction with the environment. A central complication in practice is that real-world datasets can be corrupted, by logging errors, distribution shift, or even adversarial manipulation, so algorithms must be robust to a nontrivial fraction of contaminated trajectories. In this work we study corruption-robust offline RL under *linear function approximation*, where both rewards and transitions admit linear models in a feature map.

Much of the existing offline RL theory analyses regimes with a modest number of features and shows that the sample size N must scale polynomially with the ambient dimension d to yield near-optimal policies. However, contemporary applications often operate with feature representations whose dimension d far exceeds N, an increasingly common situation with deep models, rendering these guarantees vacuous or demanding prohibitively large datasets. To address this challenge, a common strategy is to assume the model exhibits a low-dimensional structure, with sparsity being a prominent form: only a small subset $S \subset [d]$ (with $s = |S| \ll d$) significantly influences the reward and transition probability. This setting is known as the *sparse MDP* [Golowich et al., 2024] and allows the sample size to scale polynomially with the sparsity level s instead of the ambient dimension d.

Despite the importance of high-dimensionality, the sparse MDP literature, especially under corruption, remains under-explored. Existing positive results typically require strong coverage assumptions, such as *uniform coverage* [Hao et al., 2021]. By contrast, realistic data may exhibit *weak* coverage concentrated around a small set of policies. To clarify what is pos-

sible in such settings, we focus on the *single-policy concentrability* regime, where data cover only one good policy (e.g., the optimal policy), and ask:

When $d > N$ and only single-policy concentrability coverage holds, can we exploit sparsity to learn a near-optimal policy? Moreover, can we extend these guarantees to the setting where part of the dataset is corrupted?

In this paper, we tackle these questions and make significant progress in these directions, which can be summarised below.

1. **Difficulty of integrating sparsity into LSVI.** In sparse MDPs with *single-policy concentrability coverage*, directly incorporating sparsity into the LSVI framework may fail, even without corruption. The core issue is that the pointwise pessimistic bonus, central to standard LSVI analysis, is incompatible with sparsity: without support knowledge, such bonuses cause excessive Bellman error and the guarantees break down.

2. **A sparsity-aware pessimistic Actor-Critic (AC) algorithm.** To address this limitation, we develop a *pessimistic AC* framework that bypasses pointwise pessimistic bonuses. Under single-policy concentrability and no contamination, we obtain a suboptimality gap of order $\widetilde{O}(H^2 N^{-1/4}\sqrt{\kappa s})$, where H is the horizon, N the sample size, and κ the relative condition number.

3. **A sparsity-aware pessimistic AC with contamination.** We further extend our results to the contaminated setting. By integrating sparse robust regression oracles into the critic, our method achieves meaningful guarantees under both uniform coverage and single-policy concentrability, even when $d > N$ and an ε fraction of trajectories are corrupted. In particular, we obtain a suboptimality gap of $O(\sqrt{s\varepsilon})$ with a statistically optimal but computationally expensive oracle, and $O(\sqrt{s}\,\varepsilon^{1/4})$ with a computationally efficient oracle.

To the best of our knowledge, ours is the first result to achieve near-optimal policy learning in the high-dimensional regime $d > N$ under single-policy concentrability. Our findings also reveal a sharp contrast between two prominent offline RL paradigms: although both LSVI and AC methods can be near-optimal in *non-sparse* MDPs, enforcing pessimism in a pointwise manner is unnatural for LSVI analysis and renders the guarantee vacuous in the *sparse* setting, whereas the AC approach naturally accommodates sparsity and provides non-vacuous guarantees. To save space, all proofs in this paper are deferred to the Appendix.

1.1 Related works

We now briefly outline related work and refer the reader to Appendix A for a more in-depth literature review.

Offline RL. The central challenge in offline RL is to learn a near-optimal policy from data that provides only partial information about the environment. To address this challenge, value-iteration-based methods [Buckman et al., 2021, Liu et al., 2020, Kumar et al., 2019, Jin et al., 2020, Yu et al., 2020] and AC methods [Levine et al., 2020, Wu et al., 2019, Zanette et al., 2021b] are among the most popular approaches. We consider methods that apply pessimism by penalising value functions of policies that are under-represented in the data. However, as we show in Section 4.2, such an approach may lead to vacuous bounds in the sparse setting, arguably due to an overcompensation for uncertainty arising from the use of pointwise pessimistic bonuses. This also motivates the use of AC methods, which we show can effectively mitigate such an issue.

Corruption-robust offline RL. In this paper, we consider data poisoning attacks in offline RL [Zhang et al., 2021]. Previous work on corruption-robust offline RL has primarily considered the low-dimensional setting where $N \gg d$ [Zhang et al., 2021, Ye et al., 2023, Mandal et al., 2025]. In contrast, this work focuses on the high-dimensional setting $d > N$. In particular, we employ a value function estimation approach [Zhang et al., 2021] based on sparse robust estimation [Merad and Gaïffas, 2022]. We find that, under single-policy concentrability, in sparse settings, integrating sparse oracles into LSVI-type algorithms is considerably more challenging and may yield suboptimal bounds that depend polynomially on d, in contrast to prior work in the low-dimensional setting [Zhang et al., 2021]. Moreover, we demonstrate that sparse robust estimation can be naturally integrated into the AC framework, effectively eliminating polynomial dependence on d. A systematic comparison of our results with previous work on corruption-robust offline RL is provided in Table 1.

Sparse Linear MDP. Sparse linear MDPs have been a primary focus in the online RL literature [Golowich et al., 2024, Kim et al., 2024, Hao et al., 2021], where one can often design exploratory policies to ensure good data coverage and thereby exploit the underlying sparsity structure. In contrast, there are far fewer works on sparse MDPs in the offline RL setting [Hao et al., 2021], where strong data coverage assumptions (e.g., uniform coverage) are typically imposed to bypass the need for pessimism, as we shall explain later in this paper. However, we note that *limited data coverage* is the central challenge of offline RL compared to online RL, without an exploratory policy that guarantees sufficient coverage, it is unclear from prior work whether one can learn a near-optimal policy. For

	Suboptimality gap	#Sample required	Non-vacuous bound when $d > N$?	Coverage assumption
[Ye et al., 2023]	$\widetilde{O}\left(\frac{H^3\sqrt{d\epsilon}}{\xi}\right)$	poly(d, H)	✗	Uniform coverage
[Zhang et al., 2021]	$\widetilde{O}\left(\frac{H^3\sqrt{d\epsilon}}{\xi}\right)$	poly(d, H)	✗	Uniform coverage
Ours	$\widetilde{O}\left(\frac{H^3 s\sqrt{\epsilon}}{\xi}\right)$	poly(s, H)	✓	Uniform coverage
[Zhang et al., 2021]	$\widetilde{O}\left(H^3\sqrt{d\kappa\epsilon}\right)$	poly(d, H)	✗	Single-policy concentrability
Ours	$\widetilde{O}\left(H^3\sqrt{s\kappa\epsilon}\right)$	poly(s, H)	✓	Single-policy concentrability

Table 1: Comparisons with existing results on corruption-robust offline RL. The table compares our results with existing results in the literature on corruption-robust offline RL. The column *Non-vacuous bound* means that, in the clean setting, the suboptimality gap can approach 0 when $d > N$. Our result in this table is stated under the boundedness assumption $\|\phi(\cdot)\|_\infty \le 1$ and $\max_h \max\{\|\theta_h\|_1, \|\mu_h(\mathcal{X})\|_1\} \le 1$ for fair comparison. In the *Coverage assumption* column, *Uniform coverage* means that the covariance matrix has minimum eigenvalue at least ξ (see Assumption 2.2), while *Single-policy concentrability coverage* means that the relative condition number is at most κ (see Assumption 2.3).

this reason, our paper primarily focuses on the case of limited coverage, namely *single-policy concentrability*, and is the first to establish non-vacuous suboptimality guarantees for sparse MDPs in this regime.

2 PRELIMINARIES

Notations. For $K \in \mathbb{N}^+$, denote $[K] = \{1, \dots, K\}$. Let $\mathbb{I}$ be the indicator function. For $z \in \mathbb{R}^d$ and $S \subseteq [d]$, let $z_S \in \mathbb{R}^d$ be the vector obtained by zeroing out coordinates outside S, i.e. $(z_S)_i = z_i \mathbb{I}\{i \in S\}$. For any symmetric matrix $\Sigma \in \mathbb{R}^{d\times d}$ and index set $S \subseteq [d]$, let $\Sigma_S \in \mathbb{R}^{|S|\times|S|}$ denote the principal submatrix of Σ indexed by S, i.e., $\Sigma_S = \left[\Sigma_{ij}\right]_{i\in S,\, j\in S}$. For a symmetric matrix Σ, denote $\lambda_{\min}(\Sigma)$ as its smallest eigenvalue.

General MDPs. Consider an episodic MDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, H, P, R, x_1)$, where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, H is the episode length, $P = \{P_h : \mathcal{X}\times\mathcal{A} \to \Delta(\mathcal{X})\}_{h=1}^H$ denotes the transition probabilities (here, $\Delta(\mathcal{X})$ denotes the space of probability distributions over $\mathcal{X}$), R is a bounded stochastic reward with support $\mathrm{supp}(R_h) = [0, 1]$ and mean $r_h : \mathcal{X}\times\mathcal{A} \to [0, 1]$ for all $h \in [H]$, x_1 is the initial state. For a policy π, let $d^\pi = (d_h^\pi)_{h=1}^H$ denote the occupancy measures over state-action pairs induced by π and the transition probability P, that is, $d_h^\pi = \mathbb{P}(x_h = x, a_h = a \mid \pi, x_1)$.

For any function $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, we define the π-Bellman operator and the (optimal) Bellman operator

$$(\mathbb{B}_h^\pi f)(x, a) \triangleq r_h(x, a) + \mathop{\mathbb{E}}_{\substack{x'\sim P_h(\cdot|x,a)\\ a'\sim\pi(\cdot|x')}}[f(x', a')];$$

$$(\mathbb{B}_h f)(x, a) \triangleq r_h(x, a) + \mathop{\mathbb{E}}_{x'\sim P_h(\cdot|x,a)}\left[\max_{a'\in\mathcal{A}} f(x', a')\right].$$

We then have the Bellman equation

$$Q_h^\pi(x, a) = (\mathbb{B}_h^\pi Q_{h+1}^\pi)(x, a),\ V_h^\pi(x) = \mathop{\mathbb{E}}_{a\sim\pi(\cdot|x)}[Q_h^\pi(x, a)],$$

and the Bellman optimality equation

$$Q_h^\star(x, a) = (\mathbb{B}_h Q_{h+1}^\star)(x, a),\ V_h^\star(x) = \max_{a\in\mathcal{A}} Q_h^\star(x, a).$$

Let $\pi_\star$ be an optimal policy, for a policy π, define the sub-optimal gap

$$\mathrm{SubOpt}(\pi_\star, \pi) \triangleq V^{\pi_\star}(x_1) - V^\pi(x_1).$$

Assumption 2.1 (Sparse linear MDP)**.** *There exists a feature map $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, signed measures $\mu_h : \mathcal{X} \to \mathbb{R}^d$, and parameter vectors $\theta_h \in \mathbb{R}^d$ such that, for all (x', x, a),*

$$\begin{aligned} P_h(x' \mid x, a) &= \langle\phi(x, a), \mu_h(x')\rangle, \\ r_h(x, a) &= \langle\phi(x, a), \theta_h\rangle. \end{aligned} \tag{1}$$

We assume that the MDP is s-sparse, that is, there is $S \subset [d]$, $|S| = s$, such that for any $i \notin S$, $(\mu_h(x))_i = 0$ for all $x \in \mathcal{X}$, and $\theta_i = 0$. We assume that for all $(x, a, h) \in \mathcal{X} \times \mathcal{A} \times [H]$, $\|\phi(x, a)\|_\infty \le 1$, $\|\theta_h\|_1 \le 1$, and $\|\mu_h(\mathcal{X})\|_1 \le 1$.

Next, we can define the space of linear Q-function as follows: for any h, define

$$\mathcal{Q}_h = \{(x, a) \mapsto \langle\phi(x, a), w\rangle \mid w \in \mathbb{R}^d,\ \|w\|_1 \le H + 1 - h,\ \|w\|_0 \le s\}.$$

Offline data set. We consider the offline setting with contaminated data. Let $\widetilde{\mathcal{D}}$ be a clean dataset consisting of N trajectories, i.e., $\widetilde{\mathcal{D}} = \{(x_h^\tau, a_h^\tau, R_h^\tau)\}_{\tau=1,h=1}^{N,H}$. Suppose there is an adversary who observes the dataset $\widetilde{\mathcal{D}}$ and may arbitrarily corrupt up to ϵN trajectories, resulting in a corrupted dataset $\mathcal{D}$. We split the dataset into H disjoint subsets $\mathcal{D}_1, \dots, \mathcal{D}_H$, and use $\mathcal{D}_h$ only for the stage-h regression. In this work, regarding clean data set, we assume there is an exploratory policy π_ν,

with the occupancy probability $\nu = (\nu_h)_{h=1}^H$. Define $\Sigma_h \triangleq \mathbb{E}_{\nu_h}\left[\phi(x,a)\phi(x,a)^\top\right]$ as the covariance matrix of the underlying data distribution ν at horizon h. We introduce an assumption on the distribution ν, which is crucial for our main result.

Assumption 2.2 (Uniform coverage)**.** *We say that the data distribution ν is ξ-covering the state–action space with $\xi > 0$ if, for any $h \in [H]$, $\Sigma_h \succeq \xi I$, i.e., the smallest eigenvalue of Σ_h is at least ξ.*

Assumption 2.3 (Sparse single-policy concentrability)**.** *Let $\Sigma_{\star,h} = \Sigma_{\pi_\star,h}$ denote the covariance matrix induced by the optimal policy $\pi_\star$ at horizon h. For any $h \in [H]$ and vector $z \in \mathbb{R}^d$ with $\|z\|_0 \le 2s$, assume that*

$$\max_{z\in\mathbb{R}^d:\|z\|_0\le 2s} \frac{z^\top \Sigma_{\star,h} z}{z^\top \Sigma_h z} \le \kappa. \tag{2}$$

3 SPARSE ROBUST LINEAR REGRESSION

Suppose there is a collection of data points $\widetilde{\mathcal{D}} = \{(z_i, y_i)\}_{i=1}^N$ generated i.i.d. from the underlying model $y = z^\top w_\star + \eta$, where $z \sim \nu \in \Delta(\mathbb{R}^d)$ satisfies $\|z\|_\infty \le 1$. Denote the covariance matrix as $\Sigma = \mathbb{E}_{z\sim\nu}[zz^\top]$. The unknown parameter $w_\star \in \mathbb{R}^d$ is an s-sparse vector, and η is zero-mean sub-Gaussian noise with $\mathrm{Var}(\eta) \le \sigma^2$. Suppose an adversary can arbitrarily corrupt up to ϵN data points, resulting in a corrupted dataset $\mathcal{D}$, and the estimator has access only to $\mathcal{D}$. We now discuss the estimation error of different robust sparse linear estimators under various settings of Σ.

Given a dataset $\mathcal{D}$, the Sparse Robust Linear Estimator, denoted `SRLE`, is a mapping that takes $\mathcal{D}$ as input and outputs an estimator $\widehat{w}$ of the true parameter $w_\star$, i.e., $\widehat{w} = \texttt{SRLE}(\mathcal{D})$. In this section, we will investigate different types of `SRLE` under different assumptions on the covariate distribution ν.

3.1 Robust Sparse Linear Regression with Uniform Coverage

We begin by analyzing the case where the covariate matrix Σ is well-conditioned, i.e., its minimum eigenvalue is bounded away from zero. This *uniform coverage* assumption ensures that the features are sufficiently informative across all directions, which allows us to design computationally efficient and statistically robust estimators. The following result, adapted from [Merad and Gaïffas, 2022], provides guarantees for such an estimator.

Proposition 3.1. *If $\lambda_{\min}(\Sigma) \ge \xi > 0$, then with probability at least $1-\delta$, there exists a robust least squares estimator, named `SRLE1`, which returns an estimate $\widehat{w}$ satisfying*

$$\|\widehat{w} - w_\star\|_1 = O\left((\sigma + \|w_\star\|_1)\left(\frac{s\log(d/\delta)}{\xi\sqrt{N}} + \frac{s\sqrt{\epsilon}}{\xi}\right)\right). \tag{3}$$

Moreover, the estimator runs in $\mathrm{poly}(d, s, N)$ time.

To the best of our knowledge, `SRLE1` is a practical estimator that achieves a small error bound when the covariate distribution is sub-Gaussian, although we suspect its guarantee is highly suboptimal. Since developing robust estimators is not the main focus of this paper, but rather understanding how to use them effectively in offline RL, further advances in sparse regression under corruption could directly strengthen our results by replacing the `SRLE1` oracle with a stronger alternative.

Proposition 3.1 shows that, under uniform coverage, we can achieve both robustness to data contamination and computational efficiency. However, in practice the uniform coverage assumption may not hold, especially in high-dimensional settings where the covariate matrix can be ill-conditioned, even when restricted to sparse subspaces. We therefore turn to the more challenging case where no such assumption is imposed.

3.2 Robust Sparse Linear Regression without Uniform Coverage

When Σ may be ill-conditioned, that is, $\lambda_{\min}(\Sigma) \to 0$, the estimation error bound in Proposition 3.1 becomes vacuous. In this setting, we first present a statistically optimal but computationally expensive estimator, and then show another computationally efficient estimator with larger statistical error. Due to the lack of uniform coverage, the estimation error in terms of $\|\cdot\|_1$ can no longer be guaranteed. Instead, we provide estimation error bounds with respect to the Σ-norm $\|\cdot\|_\Sigma$, which will be sufficient for the RL algorithm used later.

A non-computationally efficient estimator. The following theorem establishes the existence of an estimator, denoted `SRLE2`, which achieves minimax-optimal statistical guarantees even without uniform coverage.

Proposition 3.2. *For any covariate matrix Σ, with probability at least $1-\delta$, there exists an estimator named `SRLE2` returns $\widehat{w}$ such that*

$$\|\widehat{w} - w_\star\|_\Sigma^2 = O\left(\frac{\sigma\|w_\star\|_2\sqrt{s}\log(d/\delta)}{\sqrt{N}} + \sigma^2\epsilon\right). \tag{4}$$

Note that, in the case of an ill-conditioned covariate matrix, the minimax optimal rate is $1/\sqrt{N}$ [Hsu et al., 2014], which is known as the slow rate in the literature. We also note that previous work such as [Jin et al., 2021, Zhang et al., 2021] uses an oracle with rate $1/N$, but with a hidden constant that scales inversely with $\lambda_{\min}(\Sigma)$, which is invalid in our setting as $\lambda_{\min}(\Sigma) \to 0$.

A computationally efficient estimator. While SRLE2 enjoys optimal statistical guarantees of order $O(\epsilon)$, it is computationally intractable in general (see Appendix C.2 for details), since best subset selection is a NP-hard problem. To address this limitation, one can employ an alternative algorithm, SRLE3, which is polynomial-time computable at the cost of looser statistical guarantees.

Proposition 3.3. *For any covariate matrix Σ, there exists an algorithm named SRLE3 such that, with probability at least $1-\delta$, it returns an estimate $\widehat{w}$ satisfying*

$$\|\widehat{w} - w_\star\|_\Sigma^2 = O\Bigg(\|w_\star\|_1(\|w_\star\|_1 + \sigma) \times \left(\sqrt{\frac{\log(d\delta^{-1})}{N}} + \sqrt{\epsilon} \right) \Bigg).$$

Moreover, the estimator runs in $\mathrm{poly}(N, d, s)$ *time.*

Note that the statistical error for SRLE3 is $O(\sqrt{\epsilon})$, in contrast to $O(\epsilon)$ for the computationally expensive oracle SRLE2. This trade-off highlights the fundamental tension in the absence of uniform coverage: one may either obtain statistically optimal but computationally expensive estimators, or efficient algorithms with degraded accuracy. We also note that the bound in Proposition 3.3 can have implicit dependence on s through $\|w_\star\|_1$.

3.3 Using SRLE Estimators in Offline Reinforcement Learning

So far, we have established guarantees for several variants of Sparse Robust Linear Estimators (SRLE) under different assumptions on the covariate distribution. While these regression results are of independent interest, their main role in this paper is to serve as a key component in the design and analysis of offline reinforcement learning algorithms. In particular, the SRLE estimators will be used to approximate the linear predictors that arise when estimating value functions.

Formally, recall that for each horizon h, the population covariance is

$$\Sigma_h = \mathbb{E}_{\nu_h}\left[\phi(x,a)\phi(x,a)^\top\right],$$

and its empirical analogue is defined as

$$\widehat{\Sigma}_h \triangleq \frac{1}{|\mathcal{D}_h|} \sum_{(x,a)\in\mathcal{D}_h} \phi(x,a)\phi(x,a)^\top + (\lambda+\epsilon)I.$$

Let $\mathcal{F}$ denote the space of real-valued functions on $\mathcal{X}\times\mathcal{A}$. Given a policy π, first we define a data set compatible with SRLE using the offline data set $\mathcal{D}$ and policy π, as

$$\mathcal{D}_h^\pi \triangleq \left\{ \left(\underbrace{\phi(x_h^\tau, a_h^\tau)}_{z_i}, \underbrace{R_h^\tau + \mathbb{E}_{\pi_{h+1}}\left[F(x_{h+1}^\tau, a)\right]}_{y_i} \right) \right\}_{\tau=1}^{N/H}.$$

Next, we define three operators, a projection operator $\mathcal{P}^\pi$, a regression operator $\mathcal{R}^\pi$, and the associated estimation error $\mathcal{E}^\pi$, as mappings from $\mathcal{F}$ to $\mathbb{R}^d$:

$$\begin{aligned} \mathcal{P}_h^\pi(F) &\triangleq \theta_h + \int_{\mathcal{X}} \left(\sum_{a'\in\mathcal{A}} \pi(a'|x')F(x',a') \right) \mu_h(x')\,dx', \\ \mathcal{R}_h^\pi(F) &\triangleq \texttt{SRLE}(\mathcal{D}_h^\pi), \\ \mathcal{E}_h^\pi(F) &\triangleq \mathcal{P}_h^\pi(F) - \mathcal{R}_h^\pi(F). \end{aligned} \tag{5}$$

In words, for any function F, the projection operator $\mathcal{P}_h^\pi(F)$ represents the best-fit linear predictor, the regression operator $\mathcal{R}_h^\pi(F)$ applies an SRLE estimator to approximate this predictor using corrupted data, and $\mathcal{E}_h^\pi(F)$ captures the resulting estimation error. Note that the precise statistical guarantees of $\mathcal{R}_h^\pi$ depend on the choice of SRLE variant (e.g., SRLE1, SRLE2, SRLE3), and we will specify this choice in each subsequent result.

For later use, we also define the greedy projection operator $\mathcal{P}^*$, greedy regression operator $\mathcal{R}^*$ and the associated estimation error $\mathcal{E}^*$ which replaces the policy expectation with a maximisation:

$$\begin{aligned} \mathcal{D}_h^* &\triangleq \left\{ \left(\phi(x_h^\tau, a_h^\tau),\ R_h^\tau + \max_{a\in\mathcal{A}} F(x_{h+1}^\tau, a) \right) \right\}_{\tau=1}^{N/H}. \\ \mathcal{P}_h^*(F) &\triangleq \theta_h + \int_{\mathcal{X}} \max_{a'\in\mathcal{A}} F(x',a')\mu_h(x')\,dx', \\ \mathcal{R}_h^*(F) &\triangleq \texttt{SRLE}(\mathcal{D}_h^*), \\ \mathcal{E}_h^*(F) &\triangleq \mathcal{R}_h^*(F) - \mathcal{P}_h^*(F). \end{aligned} \tag{6}$$

4 WHY ROBUST LSVI MAY FAIL IN SPARSE OFFLINE RL

In the offline RL literature, the Least Square Value Iteration (LSVI) framework is a popular choice [Jin et al., 2021]. Moreover, they can successfully handle data corruption, which makes them a strong candidate for addressing the problem studied in this paper [Zhang et al., 2021]. However, in this section we show that while LSVI-type algorithms can succeed under the uniform coverage assumption (Assumption 2.2), they can fail when only the single concentrability assumption (Assumption 2.3) is imposed. We begin by describing the Sparse Robust LSVI algorithm, whose pseudocode is provided in Algorithm 1.

Algorithm 1 Sparse Robust Least-Square Value Iteration

Input data set $\mathcal{D}$, `SRLE` oracle, Pessimistic bonus function $\{\Gamma_h\}_{h=1}^H$.
Initialise $\underline{Q}_{H+1} = \mathbf{0}$.
for $h \in [H]$ **do**
  Set $\widehat{w}_h \leftarrow \mathcal{R}_h^*(\underline{Q}_{h+1})$.
  Set $\underline{Q}_h(x,a) = \phi(x,a)^\top \widehat{w}_h + \Gamma_h(x,a)$, clipped in $[0, H+1-h]$.
  Set $\widehat{\pi}_h(a \mid x) = \begin{cases} 1 \text{ if } a = \arg\max_a \underline{Q}_h(x,a) \\ 0 \text{ otherwise .} \end{cases}$.
end for
Output $\{\widehat{\pi}_h\}_{h=1}^H$.

4.1 Sparse Robust LSVI with Uniform Coverage

The uniform coverage assumption enables the use of `SRLE1` estimator in Proposition 3.1, which in turn allows us to apply the pessimistic LSVI framework for robust offline RL.

Proposition 4.1. *Suppose Assumption 2.1 and 2.2 hold. Let the pessimistic bonus be $\Gamma_h(x,a) = 0$ for all $h \in [H]$, $x \in \mathcal{X}$, $a \in \mathcal{A}$. Run Algorithm 1 with the `SRLE1` estimator. Then, with probability at least $1-\delta$.*

$$\mathrm{SubOpt}(\pi_\star, \widehat{\pi}) = O\left(\frac{H^3 s \log(d/\delta)}{\xi\sqrt{N}} + \frac{H^3 s\sqrt{\epsilon}}{\xi}\right). \quad (7)$$

Remark 4.2. Under the uniform coverage assumption, the parameter ξ can be regarded as a constant (independent of d) in many settings, as long as the feature vectors satisfy $\|\phi(\cdot)\|_\infty \le 1$. For example, when $\phi(x,a)$ is sampled from $\{-1,1\}^d$ under the uniform distribution, we have $\xi = 1$. Therefore, the key feature of Proposition 4.1 is that the suboptimality gap contains no polynomial dependence on the ambient dimension d, making the result meaningful even in the high-dimensional regime $d > N$. Moreover, when $N \ge d$, the result of Zhang et al. [2021] achieves $O(H^2\sqrt{d}\xi^{-1}\epsilon)$, since $\|\phi(\cdot)\|_2 \le \sqrt{d}$ in our setting. The dependence on $\sqrt{d}$ makes this result less desirable compared to our result, $O(H^2 s\xi^{-1}\sqrt{\epsilon})$.

Moreover, our bound currently scales as $O(s\sqrt{\epsilon})$, which we believe may be further improved. In particular, if one could design an `SRLE` oracle satisfying $\|\widehat{w} - w_\star\|_1 = O(\sqrt{s}\epsilon)$, then a sharper $O(\sqrt{s}\epsilon)$ suboptimality rate would follow immediately. However, the best computationally-efficient oracle we are aware of, `SRLE1` (see Proposition 3.1), achieves only $\|\widehat{w} - w_\star\|_1 \le O(s\sqrt{\epsilon})$, which we suspect is suboptimal. Thus, further progress in sparse regression under corruption could directly tighten our results by replacing the `SRLE1` oracle.

4.2 Sparse Robust LSVI with Single Concentrability

Beyond the uniform coverage case, we now explain why pointwise pessimistic bonuses can lead to large error bounds. In this section, we assume Assumption 2.3 holds. Since the uniform coverage assumption no longer applies, we must replace the `SRLE1` estimator with either `SRLE2` or `SRLE3`. For concreteness, we focus on the result using `SRLE2`; the same challenges and limitations of LSVI also apply when using `SRLE3`. We begin by relating the suboptimality gap of pessimistic LSVI to an upper bound on the Bellman error.

Theorem 4.3 ([Jin et al., 2020])**.** *Choose a pessimistic bonus $(\Gamma_h)_{h=1}^H$ such that $(\underline{Q}_h - \mathbb{B}_h\underline{Q}_{h+1})(x,a) \le \Gamma_h(x,a)$ for any $(x,a) \in \mathcal{X}\times\mathcal{A}$ and $h \in [H]$. Then, the LSVI algorithm (Algorithm 1) outputs a policy $\widehat{\pi}$ such that* $\mathrm{SubOpt}(\pi, \widehat{\pi}) \le 2\sum_{h=1}^H \mathbb{E}_\pi[\Gamma_h(x,a)]$.

Let $w_h^\star = \mathcal{P}_h^*(\underline{Q}_{h+1})$ denote the best-fit linear predictor given the pessimistic estimator at stage $(h+1)$. Let $\widetilde{S}_h$ be the sparsity support of $\widehat{w}_h - w_h^\star$, so that $|\widetilde{S}_h| \le 2s$. The Bellman error and the pessimistic bonus Γ_h can then be bounded as

$$\begin{aligned} |\widehat{Q}_h(x,a) - \mathbb{B}_h(\underline{Q}_{h+1})(x,a)| &= |\phi(x,a)^\top(\widehat{w}_h - w_h^\star)| \\ &\overset{(i)}{\le} \underbrace{\|\widehat{w}_h - w_h^\star\|_{\widehat{\Sigma}_h}}_{\text{Estimator error } \alpha_h} \left\|\phi_{\widetilde{S}_h}(x,a)\right\|_{\widehat{\Sigma}_h^{-1}} \\ &\overset{(ii)}{\le} \alpha_h \left(\max_{S:\,|S|\le 2s} \|\phi_S(x,a)\|_{\widehat{\Sigma}_h^{-1}}\right) \triangleq \Gamma_h(x,a). \end{aligned} \quad (8)$$

Inequality (i) holds because only features in $\widetilde{S}_h$ of ϕ contribute non-zero terms to the Bellman error. The max operator in (ii) is necessary since the true support $\widetilde{S}_h$ is unknown. Thus, to guarantee pessimism, one must maximize over all subsets S of size at most $2s$. We note that $\max_{S:\,|S|\le 2s} \|\phi_S(x,a)\|_{\widehat{\Sigma}_h^{-1}}$ is not always smaller than $\|\phi(x,a)\|_{\widehat{\Sigma}_h^{-1}}$. However, restricting $\phi(\cdot)$ to a sparse support is a natural way to remove the explicit dependence on d in the Bellman error bound. Nevertheless, this maximization introduces additional error, since

$$\begin{aligned} &\mathop{\mathbb{E}}_{(x,a)\sim d^{\pi_\star}}[\Gamma_h(x,a)] \\ &\asymp \alpha_h \,\mathbb{E}_{d^{\pi_\star}}\left[\max_{S:|S|\le 2s} \left(\phi_S^\top(x,a)\Sigma_h\phi_S(x,a)\right)\right] \\ &\overset{(i)}{\ge} \alpha_h \max_{S:|S|\le 2s} \mathbb{E}_{d^{\pi_\star}}\left[\phi_S^\top(x,a)\Sigma_h\phi_S(x,a)\right]. \end{aligned} \quad (9)$$

where inequality (i) follows from Jensen's inequality by exchanging expectation and maximization.

The LHS of inequality (i) in (9) serves as the ideal upper bound for the Bellman error, contributing a

factor of order $\alpha_h\sqrt{s\kappa}$ under Assumption 2.3 [Zhang et al., 2021]. However, since sparsity support $\widetilde{S}_h$ is unknown, the Bellman error must be bounded by the RHS, which can be significantly larger as shown below.

Proposition 4.4. *Let $\Sigma = \lambda I$. Let $z \sim \mathrm{Ber}^d(1/2)$. Then, for $d > 4s$, we have*

$$\underbrace{\mathbb{E}_{z\sim\mu}\Big[\max_{S:|S|=2s} z_S^\top \Sigma^{-1} z_S\Big]}_{LHS\ in\ (9)} - \underbrace{\max_{S:|S|=2s} \mathbb{E}_{x\sim\mu}\big[z_S^\top \Sigma^{-1} z_S\big]}_{RHS\ in\ (9)} \geq (1 - 2\exp(-d/8))\frac{s}{\lambda}. \tag{10}$$

Since λ is choose typically of order $1/\sqrt{N}$, the resulting $s\log(d)\sqrt{N}$-multiplicative factor in the Bellman error, leading to a large suboptimality gap for LSVI.

Remark 4.5. The difficulty with LSVI is that it imposes pessimism in a pointwise manner. This can be viewed as *over-pessimism*. While such over-pessimism does not cause issues in low-dimensional MDPs, it leads to excessive Bellman error in high-dimensional sparse MDPs, where the sparsity support is hidden from the learner. Let $\underline{V}_1(x_1) = \max_a \underline{Q}_1(x_1, a)$, from [Jin et al., 2021], the main reason LSVI requires pointwise pessimistic bonus is the need to enforce

$$\underline{V}_1(x_1) - V_1^{\widehat{\pi}}(x_1) = \mathbb{E}_{d^{\widehat{\pi}}}\left[\sum_{h=1}^{H} (\underline{Q}_h - \mathbb{B}_h \underline{Q}_{h+1})(x, a)\right] < 0.$$

Here, the policy $\widehat{\pi}$ is defined greedily with respect to $(\underline{Q}_h)_{h=1}^H$, which means that $\widehat{\pi}$ is chosen *after* $(\underline{Q}_h)_{h=1}^H$ is fixed. This creates a statistical dependency: $\widehat{\pi}$ tends to place weight on (x, a) where $(\underline{Q}_h - \mathbb{B}_h \underline{Q}_{h+1})(x, a)$ is less negative (or even positive), due to the greedy nature of $\widehat{\pi}$ (i.e., the max operator). As a result, the only distribution-free way to guarantee the inequality above is to impose

$$(\underline{Q}_h - \mathbb{B}_h \underline{Q}_{h+1})(x, a) \leq 0 \quad \forall (x, a, h), \tag{11}$$

i.e., using pointwise pessimistic bonus. We shall see that the AC method introduced in Section 5.2 avoids this requirement and can achieve a meaningful suboptimality gap as a result.

5 SPARSE ACTOR-CRITIC METHODS

The shortcoming of LSVI arises from its pointwise pessimistic bonus. In this section, we show that this issue can be addressed by incorporating sparsity directly into AC methods. The key idea is that, unlike LSVI, AC does not impose pessimism uniformly over all state–action pairs; instead, the critic only needs to evaluate the current actor's policy pessimistically. This is better aligned with sparse structure, since the regression error is controlled only along the policy being optimized, thereby avoiding the overly conservative bonus that limits LSVI. As a result, sparsity, pessimism, and weak coverage can be combined in a natural way.

First, consider the actor. We use a log-linear class of policies. In particular, for any policy parameter $\upsilon = (\upsilon_h)_{h=1}^H$, we define

$$\pi_h^\upsilon(a \mid x) = \frac{\exp(\langle \phi(x, a), \upsilon_h \rangle)}{\exp\big(\sum_{a\in\mathcal{A}} \langle \phi(x, a), \upsilon_h \rangle\big)}.$$

To update the actor, we invoke the mirror descent framework [Zanette et al., 2021b], that is, $\pi_{t+1,h} \propto \pi_{t,h} \exp(\eta Q_{h,t}(x, a))$. When the Q-function is linear, that is, $Q_{h,t} \in \mathcal{Q}_h$, it is possible to simplify the actor update as shown in Line 5 of Algorithm 2. The pseudocode of Sparse Robust Actor-Critic is provided in Algorithm 2.

Algorithm 2 Sparse Robust Actor-Critic

Actor:

1: Input dataset $\mathcal{D}$, learning rate η, `SRLE` oracle.
2: Initialize $\upsilon_1 = \mathbf{0}_d$.
3: **for** $t \in [T]$ **do**
4: $\quad \underline{w}_t = \texttt{Critic}(\mathcal{D}, \pi_{\upsilon_t})$.
5: $\quad \upsilon_{t+1} = \upsilon_t + \eta \underline{w}_t$
6: **end for**

Critic:

1: Input dataset $\mathcal{D}$, `SRLE` oracle.
2: Solve `PessOpt` subroutine, obtain weight $\underline{w}$.
3: Return pessimistic weight $\underline{w}$.

5.1 Sparse Robust AC with Uniform Coverage

We first consider the case where the uniform coverage assumption holds. Here, we can directly apply the `SRLE1` estimator to define the regression operator $\mathcal{R}^\pi$. This leads to a natural construction of the critic, which recursively estimates value functions in a pessimistic manner.

Formally, the subroutine `PessOpt` within the critic is defined by the following pessimistic optimisation problem. Let $\underline{Q}_{H+1} = \mathbf{0}$. Given a policy π, for $h \in [H]$, we define recursively

PessOpt (uniform coverage)**.**

$$\begin{aligned} \underline{w}_h \in \arg\min_{w_h} \quad & \left\| w_h - \mathcal{R}_h^\pi(\underline{Q}_{h+1}) \right\|_1; \\ s.t. \quad & \|w_h\|_1 \leq H + 1 - h. \end{aligned} \tag{12}$$

The overall algorithm alternates between actor and critic updates, summarized in Algorithms 2. The next

result shows that this sparse robust actor-critic algorithm achieves strong suboptimality guarantees under uniform coverage.

Theorem 5.1. *Suppose Assumption 2.1 and 2.2 hold. Then, with step size* $\eta = \sqrt{\frac{\log \mathcal{A}}{N}}$, *after* $T = N/H$ *iterations the actor-critic algorithm returns a policy* $\widehat{\pi}$ *that satisfies*

$$\mathrm{SubOpt}(\pi_\star, \widehat{\pi}) = O\Bigg(\frac{H^3 s \log(dNH\delta^{-1})}{\xi\sqrt{N}} + H^3\sqrt{\frac{\log(|\mathcal{A}|)}{N}} + \frac{H^3 s\sqrt{\epsilon}}{\xi}\Bigg), \tag{13}$$

with probability at least $1-\delta$.

This result demonstrates that, under uniform coverage, actor-critic methods achieve near-optimal sample complexity in the sparse setting, similar to LSVI-based algorithms. Moreover, we will show that actor-critic methods are superior to LSVI when the data distribution is ill-conditioned.

5.2 Sparse Robust AC with Single Concentrability

We now turn to the more challenging setting where only the single-policy concentrability assumption holds. In this case, the critic must be modified to incorporate explicit pessimistic constraints. The actor-critic algorithm remains the same in structure, but the critic subproblem is more involved.

Formally, the subroutine `PessOpt` within the critic is defined by the following optimisation problem.

PessOpt (single concentrability)**.**

$$\begin{aligned}\{\underline{w}_h^\pi\}_{h=1}^H \triangleq \underset{\{w_h\}_{h=1}^H}{\arg\min} \quad & \sum_a \pi_1(a|x_1)\langle\phi(x_1,a), w_1\rangle \\ s.t. \quad & \|w_h - \mathcal{R}_h^\pi(\langle\phi, w_{h+1}\rangle)\|_{\widehat{\Sigma}_h}^2 \le \alpha_h^2, \\ & \|w_h\|_1 \le H+1-h, \\ & \|w_h\|_0 \le s.\end{aligned} \tag{14}$$

Remark 5.2 (How AC avoids pointwise pessimistic bonuses)**.** Notice that the objective function of the `PessOpt` subroutine only ensures that the value function is pessimistic at the initial state, that is, $\underline{V}_1^\pi(x_1) \le V_1^\pi(x_1)$ for the given policy π. It does not guarantee that the value function is pessimistic for any other (x,a,h), which is the key difference compared to how LSVI imposes pointwise bonuses (8). We note that AC is able to bypass pointwise bonuses thanks to its algorithmic structure. In particular, at each phase t, the AC algorithm *fixes a policy* π_t *first*; then it is possible to compute a pessimistic Q function that defines $\mathcal{E}_h^\pi(x,a)$, which satisfies $\underline{V}_1^{\pi_t}(x_1) - V_1^{\pi_t}(x_1) < 0$, by solving the `PessOpt` subroutine (14). In contrast, in LSVI-type algorithms the policy $\widehat{\pi}$ is defined greedily *only after* $\underline{Q}_h(x,a)$ is constructed. Consequently, to safely guarantee $\underline{V}_1(x_1) - V^{\widehat{\pi}}(x_1) < 0$, one must impose pessimism in the pointwise manner as in (11).

This simple observation allows us to derive the suboptimality-gap guarantee stated below.

Theorem 5.3. *Suppose Assumption 2.1 and 2.3 hold. Let operators* $(\mathcal{R}_h^\pi)_{h=1}^H$ *be defined using the* `SRLE2` *estimator. Choose the sequence* α *such that*

$$\alpha_h^2 = O\Bigg(\frac{\sqrt{s}H^3\log(dH\delta^{-1})}{\sqrt{N}} + H^3\epsilon + H^3(\lambda+\epsilon)\Bigg).$$

Choose $\lambda = \frac{s}{N}\log\frac{d}{s\delta}$. *Then, after* $T = N$ *iterations with step size* $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{H^2N}}$, *the actor-critic algorithm returns a policy* $\widehat{\pi}$ *that satisfies*

$$\mathrm{SubOpt}(\pi_\star, \widehat{\pi}) = O\Bigg(\frac{H^3\sqrt{\kappa}s^{3/4}\sqrt{\log(dHN\delta^{-1})}}{N^{\frac{1}{4}}} + H^3\sqrt{\frac{\log(|\mathcal{A}|)}{N}} + H^3\sqrt{\kappa s\epsilon}\Bigg),$$

with probability at least $1-\delta$.

Theorem 5.3 shows that, even under weaker coverage conditions, sparse actor-critic algorithms can still achieve meaningful robustness guarantees, though with slower rates compared to the uniform coverage case. In particular, suppose $N = \Omega\Big(\frac{s\log^2(dHN\delta^{-1})}{\epsilon^2}\Big)$, then we obtain

$$\mathrm{SubOpt}(\pi_\star, \widehat{\pi}) = \widetilde{O}(H^3\sqrt{\kappa s\epsilon}). \tag{15}$$

We note that the result in Theorem 5.3 relies on the `SRLE2` oracle, which is computationally expensive. Therefore, we derive the next theorem, which leverages the more efficient `SRLE3` oracle, at the cost of a larger suboptimality gap.

Theorem 5.4. *Suppose Assumption 2.1 and 2.3 hold. Let operator* $\mathcal{R}_h^\pi$ *be defined using the* `SRLE3` *estimator. Choose the sequence* α *such that*

$$\alpha_h = O\Bigg(\frac{H^3\log(dH\delta^{-1})}{\sqrt{N}} + H^3\sqrt{\epsilon} + H^3(\lambda+\epsilon)\Bigg).$$

Choose $\lambda = \frac{s}{N}\log\frac{d}{s\delta}$. *After* $T = N$ *iterations with step size* $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{H^2N}}$, *the actor-critic algorithm returns a*

policy $\widehat{\pi}$ *that satisfies*

$$\text{SubOpt}(\pi_\star, \widehat{\pi}) = O\Bigg(\frac{H^3\sqrt{\kappa s}\sqrt{\log(dHN\delta^{-1})}}{N^{\frac{1}{4}}} + H^3\sqrt{\frac{\log(|\mathcal{A}|)}{N}} + H^3\sqrt{\kappa s}\epsilon^{\frac{1}{4}} \Bigg), \tag{16}$$

with probability at least $1-\delta$.

Theorem 5.4 shows that using computationally efficient oracle SRLE3, sparse actor-critic algorithms still achieve meaningful robustness guarantees, although slightly worse compared to that of Theorem 5.3. In particular, suppose $N = \Omega\Big(\frac{\log^2(dHN\delta^{-1})}{\epsilon}\Big)$, then we obtain

$$\text{SubOpt}(\pi_\star, \widehat{\pi}) = \widetilde{O}(H^3\sqrt{\kappa s}\epsilon^{\frac{1}{4}}). \tag{17}$$

Proof sketch of Theorem 5.3 and 5.4. The proof has three main steps. First, for each actor iterate π_t, we solve the constrained critic problem in (14) and use its solution to define an induced MDP M_t, similar to that in Zanette et al. [2021b]. By construction, the critic is exact in M_t, and the induced reward perturbation is chosen so that $V^{\pi_t}_{M_t}$ is no larger than the true value V^{π_t}. Thus, the critic provides a pessimistic evaluation of the current policy without requiring a pointwise lower bound over all state-action pairs.

Second, the sparse robust regression oracle controls the critic error in the empirical covariance norm. The key step is then to transfer this control from the data distribution to the occupancy measure of the comparator policy. Under single-policy concentrability, this transfer incurs only a $\sqrt{\kappa}$ factor, while sparsity ensures that the complexity depends on the support size s, rather than the ambient dimension d. This yields a bound on the suboptimality of evaluating $\pi_\star$ in the induced MDP.

Third, we combine this critic bound with the standard mirror-descent regret guarantee for the actor updates. Averaging over iterations and summing the two contributions gives the final suboptimality bound. In particular, the result follows from the interaction of three ingredients: pessimistic policy evaluation through the induced MDP, sparse regression error control under weak coverage, and no-regret policy optimization. □

In the uniform coverage case discussed in Sections 5.1 and 4.1, both Algorithm 1 and Algorithm 2 can be implemented in $\text{poly}(N, d, H, s)$ time. In contrast, in the single concentrability setting, a necessary condition for computational efficiency is to use the SRLE3 estimator in place of SRLE2. However, even with SRLE3, the Actor–Critic method still requires solving Equation (14), which involves an ℓ_0-type constraint. This constraint substantially increases the computational complexity, making Equation (14) the primary bottleneck in achieving a polynomial-time algorithm.

We note that dropping the ℓ_0-constraint in (14) may cause significant errors. Without restricting $\|\phi\|_{\Sigma_h^{-1}}$ to a sparsity support, the suboptimality gap incurs a multiplicative factor $\sqrt{\kappa d}$, since $\mathbb{E}_{d^{\pi_\star}}\Big[\|\phi\|_{\Sigma_h^{-1}}\Big] = O(\sqrt{\kappa d})$ in the worst case [Zhang et al., 2021]. Moreover, without additional assumptions on the data distribution ν, it remains unclear how to design an algorithm that solves Equation (14) in $\text{poly}(N, d, s)$ time. We conjecture that some relaxation of data distribution ν may be needed to make such a solution computationally feasible.

6 CONCLUSION

We studied offline RL in sparse linear MDPs with limited coverage and adversarial corruption, showing that pessimistic LSVI can fail in sparse regimes due to overly conservative pointwise pessimistic bonuses. To address this, we proposed a pessimistic actor–critic framework with sparse, robust regression oracles, achieving near-optimal guarantees when $d \gg N$ under single-policy concentrability. Our results provide the first non-vacuous bounds for sparse offline RL, establishing a separation between LSVI and AC methods.

An important direction for future work is to explore algorithmic relaxations of the ℓ_0-constraint in Equation (14) that preserve statistical guarantees while reducing computational burden. In particular, investigating distributional assumptions on ν or alternative convex surrogates could pave the way toward polynomial-time solutions.

It would also be interesting to study multi-agent Markov games under sparsity. The structure of sparse linear MDPs can be naturally extended to multi-agent Markov games. However, existing algorithms for solving offline Markov games are based on value iteration based methods, and handling sparsity will require generalization of our actor-critic framework to multi-agent setting. Furthermore, recent work Nika et al. [2024] on corruption robustness in offline Markov games has highlighted that data coverage assumptions play an important role on achievable error rates, and it would be interesting to explore how sparsity influences such data coverage assumptions.

Acknowledgements

The work of Andi Nika and Goran Radanovic was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 102–115, 2021.

Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. In *International Conference on Learning Representations*, 2021.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Exploring and learning in sparse linear mdps without computationally intractable oracles. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, 2024.

Botao Hao, Yaqi Duan, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Sparse feature selection makes batch reinforcement learning more sample efficient. In *International Conference on Machine Learning*, 2021.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14, 2014.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial Attacks on Neural Network Policies. *CoRR*, abs/1702.02284, 2017.

Natasha Jaques et al. Way Off-policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog. *CoRR*, abs/1907.00456, 2019.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems 33*, 2020.

Wonyoung Kim, Garud Iyengar, and Assaf Zeevi. A doubly robust approach to sparse reinforcement learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, 2024.

Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, 2018.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in neural information processing systems*, 32, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2020.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*, pages 45–73. Springer, 2012.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe Policy Improvement with Baseline Bootstrapping. In *International Conference on Machine Learning*, 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.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of Adversarial Attack on Deep Reinforcement Learning Agents. In *International Joint Conference on Artificial Intelligence*, 2017.

Shijie Liu, Andrew C Cullen, Paul Montague, Sarah Erfani, and Benjamin IP Rubinstein. Multi-level certified defense against poisoning attacks in offline reinforcement learning. *arXiv preprint arXiv:2505.20621*, 2025.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 2020.

Debmalya Mandal, Andi Nika, Parameswaran Kamalaruban, Adish Singla, and Goran Radanovic. Corruption robust offline reinforcement learning with human feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 4429–4437. PMLR, 2025.

Ibrahim Merad and Stéphane Gaïffas. Robust methods for high-dimensional linear learning. *Journal of Machine Learning Research*, 2022.

Andi Nika, Debmalya Mandal, Adish Singla, and Goran Radanovic. Corruption-robust offline two-player zero-sum markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251. PMLR, 2024.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy Teaching in Reinforcement Learning via Environment Poisoning Attacks. *Journal of Machine Learning Research*, 2021.

Jianwen Sun, Tianwei Zhang, Xiaofei Xie, Lei Ma, Yan Zheng, Kangjie Chen, and Yang Liu. Stealthy and Efficient Adversarial Attacks Against Deep Reinforcement Learning. In *Association for the Advancement of Artificial Intelligence*, 2020.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent Pessimism for Offline Reinforcement Learning. In *Advances in neural information processing systems*, 2021.

Chenlu Ye, Rui Yang, Quanquan Gu, and Tong Zhang. Corruption-robust offline reinforcement learning with general function approximation. In *Advances in Neural Information Processing Systems*, 2023.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 2020.

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Proceedings of Thirty Fourth Conference on Learning Theory*, 2021a.

Andrea Zanette, Martin Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021b.

Xuezhou Zhang, Yiding Chen, Jerry Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Sparse Offline Reinforcement Learning with Corruption Robustness

Contents of Appendix

Sparse Offline Reinforcement Learning with Corruption Robustness

A Related Works

Offline RL. In recent years, offline RL [Lange et al., 2012, Levine et al., 2020] has garnered considerable attention, as it offers a compelling alternative to the constraints of online data collection, making it particularly attractive for critical applications. Substantial progress has been made in this area, both on the empirical [Laroche et al., 2019, Jaques et al., 2019, Kumar et al., 2020, Kidambi et al., 2020] and theoretical fronts [Xie et al., 2021, Jin et al., 2020]. The central challenge in offline RL is learning an optimal policy from data that provides only partial information about the environment. To address this challenge, value-iteration-based methods [Buckman et al., 2021, Liu et al., 2020, Kumar et al., 2019, Jin et al., 2020, Yu et al., 2020] and actor-critic methods [Levine et al., 2020, Wu et al., 2019, Zanette et al., 2021b] are among the most popular approaches. We consider methods that apply pessimism by penalizing value functions of policies that are under-represented in the data. However, as we show in Section 4.2, such an approach may lead to vacuous bounds in the sparse setting, arguably due to an overcompensation for uncertainty arising from the use of pointwise pessimistic bonuses. This also motivates the use of actor-critic methods, which we show to effectively mitigate such an issue.

Corruption-robust offline RL. There exists a large body of literature on adversarial attacks in RL [Sun et al., 2020, Lin et al., 2017, Huang et al., 2017, Rakhsha et al., 2021], including training-time attacks, test-time attacks, and backdoor attacks. In this paper, we consider a special type of training-time attacks, namely data poisoning attacks in offline RL [Zhang et al., 2021]. To defend against such attacks, several approaches have been proposed, such as robust value function estimation based on robust statistical approaches [Zhang et al., 2021], uncertainty weighting-based methods [Ye et al., 2023] and, more recently, differential privacy-based methods [Liu et al., 2025]. In this work, we employ a value function estimation approach [Zhang et al., 2021] based on sparse robust estimation [Merad and Gaïffas, 2022]. Under uniform coverage, the use of robust linear regression oracles [Bakshi and Prasad, 2021, Klivans et al., 2018] in offline RL is straightforward. However, under weaker notions of coverage in sparse settings, we show that integrating sparse oracles into LSVI-type algorithms, a standard approach for robust offline RL, is much more challenging and may yield suboptimal bounds that depend on the full dimension, whereas dependence only on the sparse dimension would be desirable. In contrast, we demonstrate that sparse robust estimation can be naturally integrated into the actor–critic framework, effectively removing full-dimensional dependence and achieving bounds that scale only with the sparse dimension.

Sparse Linear MDP. Sparse linear MDPs have been a primary focus in the online RL literature [Golowich et al., 2024, Kim et al., 2024, Hao et al., 2021], where one can often design exploratory policies to ensure good data coverage and thereby exploit the underlying sparsity structure. In contrast, there are far fewer works on sparse MDPs in the offline RL setting [Hao et al., 2021], where strong data coverage assumptions (e.g., uniform coverage) are typically imposed to bypass the need for pessimism, as we shall explain later in this paper. However, we note that *limited data coverage* is the central challenge of offline RL compared to online RL, without an exploratory policy that guarantees sufficient coverage, it is unclear from prior work whether one can learn a near-optimal policy. For this reason, our paper primarily focuses on the case of limited coverage, namely *single-policy concentrability*, and is the first to establish non-vacuous suboptimality guarantees for sparse MDPs in this regime.

B Properties of Sparse Linear MDPs

Claim B.1. *For any $Q_{h+1} \in \mathcal{Q}_{h+1}$ and policy π, let $w_h^\pi \triangleq \mathcal{P}_h^\pi(Q_{h+1})$, then*

$$\|w_h^\pi\|_0 \le s, \quad \|w_h^\pi\|_1 \le (H-h+1).$$

Proof. By construction, we have that,

$$\begin{aligned}\|w_h^\pi\|_1 &= \left\|\theta_h + \int_{\mathcal{X}} \left(\sum_{a'\in\mathcal{A}} \pi(a' \mid x') Q_{h+1}(x', a')\right) \mu_h(x') dx'\right\|_1 \\ &\le \|\theta_h\|_1 + \|(H-h)\mu_h(\mathcal{X})\|_1 \\ &\le H - h + 1.\end{aligned} \tag{18}$$

Moreover, both θ_h and $\mu_h(x)$ are sparse vector supported in S, therefore, w_h^π has support in S, and $\|w_h^\pi\|_0 \le s$. □

Claim B.2. *For any $Q_{h+1} \in \mathcal{Q}_{h+1}$ and policy π, for any $(x', a') \in \mathcal{X} \times \mathcal{A}$, we have that*

$$\begin{aligned}\left|R_h + \sum_{a'\in\mathcal{A}} \pi(a' \mid x') Q_{h+1}(x', a')\right| &\le H - h + 1 \le 2(H-h) \\ \left|R_h + \max_{a'\in\mathcal{A}} Q_{h+1}(x', a')\right| &\le H - h + 1 \le 2(H-h).\end{aligned} \tag{19}$$

Proof. The proof is immediately followed from the fact that $|R_h| \le 1$ and $|Q_{h+1}(x', a')| \le (H-h)$ for any pair (x', a'). □

C Proof of Section 3

C.1 Estimation Error of SRLE1

SRLE1 was proposed in [Merad and Gaïffas, 2022]. It solves the convex optimization problem

$$\begin{aligned}&\min_{w\in\mathbb{R}^d} \|Xw - Y\|_2^2 \\ &\text{s.t. } \|w\|_1 \le \|w_\star\|_1,\end{aligned} \tag{20}$$

and achieves fast rates by exploiting the uniform coverage assumption via a multi-stage mirror descent algorithm. For further details, we refer the reader to the original paper [Merad and Gaïffas, 2022]. Below, we state the error guarantee of this estimator for our setting.

Proposition 3.1. *If $\lambda_{\min}(\Sigma) \ge \xi > 0$, then with probability at least $1-\delta$, there exists a robust least squares estimator, named **SRLE1**, which returns an estimate $\widehat{w}$ satisfying*

$$\|\widehat{w} - w_\star\|_1 = O\left((\sigma + \|w_\star\|_1)\left(\frac{s\log(d/\delta)}{\xi\sqrt{N}} + \frac{s\sqrt{\epsilon}}{\xi}\right)\right). \tag{3}$$

Moreover, the estimator runs in $\mathrm{poly}(d, s, N)$ *time.*

Proof. The proof is a result of Corollary 9 in [Merad and Gaïffas, 2022], which can be stated as follows. Let

$$\sigma_{\max}^2 \triangleq \max_{w:\|w\|_1\le\|w_\star\|_1} \max_{j\in[d]} \mathbb{E}_{z,y}\left[((z^\top w - y)x_j)^2\right].$$

Then, the excessive risk is bounded as

$$\|\widehat{w} - w_\star\|_1 \le O\left(\frac{\sigma_{\max} s\log(d\delta^{-1})}{\xi\sqrt{N}} + \frac{\sigma_{\max} s\sqrt{\epsilon}}{\xi}\right). \tag{21}$$

We now compute the constant $\sigma_{\max}$. Note that, since $\|z\|_\infty \le 1$, then

$$\sigma_{\max}^2 \le \max_w \mathbb{E}_{z,y}\left[(z^\top w - y)^2\right] \le \max_w \max_z \mathbb{E}_{z,y}\left[(z^\top(w - w_\star))^2\right] + \mathbb{E}[\eta^2] = 4\|w_\star\|_1^2 + \sigma^2.$$

Plug this upper bound for $\sigma_{\max}$ into the above excessive risk, we conclude the proof. □

C.2 Estimation Error of SRLE2

In this section, we will present the proof of Proposition 3.2. We first introduce the SRLE2 estimator. For any $C \subset [N]$, let $\widehat{\Sigma}(C) \triangleq \frac{1}{|C|}\sum_{i\in C} x_i x_i^\top$. Consider $\ell_0 - \ell_2$ minimisation with trimmed mean square.

$$\begin{aligned} \min_{C\subset[N]} \min_{w} \quad & \frac{1}{N}\|Y_C - Z_C w\|_2^2 + \lambda\|w\|_2^2 \\ \text{s.t.} \quad & \|w\|_0 \le s. \\ & \|w\|_1 \le \|w_\star\|_1. \\ & |C| = (1-\epsilon)n \end{aligned} \tag{22}$$

Where $Z_C = [z_i]_{i\in C}$ and $Y_C = [y_i]_{i\in C}$. We first have the upper bound for prediction error.

Lemma C.1. *Let C_* be the index set of clean data and $\widehat{C}$ be the index set returned by optimisation problem (22). For all $S \in [d]$ such that $|S| = s$, we have that*

$$\|\widehat{w} - w_\star\|^2_{\widehat{\Sigma}(\widehat{C}\cap C_*)+\lambda I} = O\left(\frac{\sigma^2 s}{N\lambda} + \frac{\sigma^2 s\log(d\delta^{-1})}{N} + 2\lambda\|w_\star\|_2^2 + \sigma^2\epsilon\right) \tag{23}$$

Proof. Based on the optimisation criteria, one has that

$$\frac{1}{N}\left\|Y_{\widehat{C}} - Z_{\widehat{C}}\widehat{w}\right\|_2^2 + \lambda\|\widehat{w}\|_2^2 \le \frac{1}{N}\|Y_{C_*} - Z_{C_*}w_\star\|_2^2 + \lambda\|w_\star\|_2^2 \tag{24}$$

This leads to

$$\begin{aligned} & \frac{1}{N}\left\|Y_{\widehat{C}} - Z_{\widehat{C}}\widehat{w}\right\|_2^2 + \lambda\|\widehat{w}\|_2^2 && \le \frac{1}{N}\|Y_{C_*} - Z_{C_*}w_\star\|_2^2 + \lambda\|w_\star\|_2^2 \\ \implies & \frac{1}{N}\left\|Y_{\widehat{C}\cap C_*} - Z_{\widehat{C}\cap C_*}\widehat{w}\right\|_2^2 + \lambda\|\widehat{w}\|_2^2 && \le \frac{1}{N}\|Y_{C_*} - Z_{C_*}w_\star\|_2^2 + \lambda\|w_\star\|_2^2 \\ \iff & \frac{1}{N}\left\|Y_{\widehat{C}\cap C_*} - Z_{\widehat{C}\cap C_*}\widehat{w}\right\|_2^2 + \lambda\|\widehat{w}\|_2^2 + \lambda\|w_\star\|_2^2 && \le \frac{1}{N}\left\|Y_{C_*\cap\widehat{C}} - Z_{C_*\cap\widehat{C}}w_\star\right\|_2^2 \\ & && \quad + \frac{1}{N}\left\|Y_{C_*\setminus\widehat{C}} - Z_{C_*\setminus\widehat{C}}w_\star\right\|_2^2 + 2\lambda\|w_\star\|_2^2 \\ \implies & \frac{1}{N}\left\|Y_{C_*\cap\widehat{C}} - Z_{C_*\cap C}\widehat{w}\right\|_2^2 + \lambda\|\widehat{w} - w_\star\|_2^2 && \le \frac{1}{N}\left\|Y_{C_*\setminus\widehat{C}} - Z_{C_*\setminus\widehat{C}}w_\star\right\|_2^2 \\ & && \quad + \frac{1}{N}\left\|Y_{C_*\cap\widehat{C}} - Z_{C_*\cap\widehat{C}}w_\star\right\|_2^2 + 2\lambda\|w_\star\|_2^2. \end{aligned}$$

Since $Y_{\widehat{C}\cap C_\star} = Z_{\widehat{C}\cap C_\star}w_\star + \eta_{\widehat{C}\cap C_\star}$, we have that $\|Y_{\widehat{C}\cap C_\star} - Z_{\widehat{C}\cap C_\star}\widehat{w}\|_2^2 = \|Z_{\widehat{C}\cap C_\star}(w_\star - \widehat{w}) + \eta_{\widehat{C}\cap C_\star}\|_2^2$. Therefore, the inequality above can be written as

$$\frac{1}{N}\left\|Z_{C_*\cap\widehat{C}}(\widehat{w} - w_\star)\right\|_2^2 + \lambda\|\widehat{w} - w_\star\|_2^2 \le \underbrace{\frac{2}{N}\left\langle \eta_{C_*\cap\widehat{C}}, Z_{C_*\cap\widehat{C}}(\widehat{w} - w_\star)\right\rangle}_{(1)} + \underbrace{\frac{1}{N}\left\|Y_{C_*\setminus\widehat{C}} - Z_{C_*\setminus\widehat{C}}w_\star\right\|_2^2}_{(2)} + 2\lambda\|w_\star\|_2^2 \tag{25}$$

Consider (1), and note that $C_* \cap \widehat{C}$ is a uncorrupted data set and $(w_\star - \widehat{w}) \in \widetilde{S}$ such that $\dim(\widetilde{S}) \le 2s$.

Claim C.2. *With probability $1-\delta$, we have that*

$$2\left\langle Z_{C_*\cap\widehat{C}}(\widehat{w} - w_\star), \eta_{C\cap C_*}\right\rangle \le \frac{1}{2}\left(\|Z_{C_*\cap C}(w_\star - \widehat{w})\|_2^2 + N\lambda\|\widehat{w} - w_\star\|_2^2\right) + O\left(\frac{\sigma^2 s}{\lambda} + \sigma^2 s\log(d\delta^{-1})\right). \tag{26}$$

Proof. We drop the subscript for data set C and subspace S to reduce clutter. Note that, $w_\star - \widehat{w}$ have support S

such that $|S| \le 2s$.

$$\begin{aligned} 2\left\langle (\widehat{w}-w_\star), Z_S^\top \eta \right\rangle &= 2\left\langle (Z_S^\top Z_S + N\lambda I)^{1/2}(\widehat{w}-w_\star), (Z_S^\top Z_S + N\lambda I)^{-1/2} Z_S^\top \eta \right\rangle \\ &\le \frac{1}{2}\left(\|Z_S(w_\star - \widehat{w})\|_2^2 + N\lambda \|(w_\star - \widehat{w})\|_2^2\right) + 2\left(\eta^\top Z_S (Z_S^\top Z_S + N\lambda I)^{-1} Z_S^\top \eta\right). \end{aligned} \tag{27}$$

Let $P_\lambda = Z_S(Z_S^\top Z_S + N\lambda I)^{-1} Z_S^\top$, we have that

$$\|P_\lambda\|_2 \le \frac{1}{\lambda}; \quad \|P_\lambda\|_F \le \frac{\sqrt{s}}{\lambda}. \tag{28}$$

Applying Hanson-Wright inequality and take union bound for all subset S such that $|S| \le 2s$, we have that with probability $1-\delta$,

$$\begin{aligned} \eta^\top P_\lambda \eta &= \mathbb{E}[\eta^\top P_\lambda \eta] + O\left(\sigma^2 s \log(d\delta^{-1})\right) \\ &= O\left(\frac{\sigma^2 \mathrm{Tr}(\widehat{\Sigma})}{\lambda} + \sigma^2 s \log(d\delta^{-1})\right) \\ &= O\left(\frac{\sigma^2 s}{\lambda} + \sigma^2 s \log(d\delta^{-1})\right). \qquad (\|z\|_\infty \le 1) \end{aligned}$$

□

Therefore,

$$\frac{2}{N}\left\langle Z_{C_*\cap C}(\widehat{w}-w_\star), \eta_{C\cap C_*}\right\rangle \le \frac{1}{2N}\|Z_{C_*\cap C}(w_\star - \widehat{w})\|_2^2 + O\left(\frac{\sigma^2 s}{N\lambda} + \frac{\sigma^2 s \log(d\delta^{-1})}{N}\right).$$

Consider (2)

$$\frac{1}{N}\left\|Y_{C_*\setminus\widehat{C}} - Z_{C_*\setminus\widehat{C}} w_\star\right\|_2^2 = \frac{1}{N}\left\|\eta_{C_*\setminus\widehat{C}}\right\|_2^2 \le 2\sigma^2\epsilon.$$

Therefore, we have that

$$\frac{1}{2N}\left\|Z_{C_*\cap\widehat{C}}(\widehat{w}-w_\star)\right\|_2^2 + \frac{\lambda}{2}\|\widehat{w}-w_\star\|_2^2 \le O\left(\frac{\sigma^2 s}{N\lambda} + \frac{\sigma^2 s\log(d\delta^{-1})}{N} + 2\lambda\|w_\star\|_2^2 + \sigma^2\epsilon\right). \tag{29}$$

As $|C_\star \cap C| \ge (1-2\epsilon)$, and assume that $\epsilon \le 1/4$, we have that $\frac{1-2\epsilon}{2(1-\epsilon)} \ge \frac{1}{4}$, $\frac{2}{1-\epsilon} \le 4$, $\frac{1}{1-\epsilon} \le 2$. Therefore,

$$\|\widehat{w}-w_\star\|_{\widehat{\Sigma}_\lambda(\widehat{C}\cap C_*)}^2 \le O\left(\frac{\sigma^2 s}{N\lambda} + \frac{\sigma^2 s\log(d\delta^{-1})}{N} + 2\lambda\|w_\star\|_2^2 + \sigma^2\epsilon\right). \tag{30}$$

□

Let $\Delta_{S,\lambda} = (\widehat{\Sigma}_{SS,\lambda})^{-1/2}(\Sigma_{SS} - \widehat{\Sigma}_{SS})(\widehat{\Sigma}_{SS,\lambda})^{-1/2}$.

Lemma C.3. *Let $\widehat{w}-w_\star$ has non-zero support in S. Assume $\|\Delta_{S,\lambda}\| \le 1$, then,*

$$\|\widehat{w}-w_\star\|_\Sigma^2 \le \|\widehat{w}-w_\star\|_{\Sigma_\lambda}^2 \le \frac{1}{1-\|\Delta_{S,\lambda}\|_2}\|\widehat{w}-w_\star\|_{\widehat{\Sigma}_\lambda}^2. \tag{31}$$

Proof. It is straightforward from the fact that $(\widehat{w}-w_\star)$ is $2s$-sparse vector, and for $\|v\|_2 \le 1$, we have that $\frac{v^\top A v}{v^\top \sigma v} \le \|\sigma^{-1}A\|_2$ if A, σ are both strictly symmetric PSD matrix. And By Lemma 3 of [Hsu et al., 2014], we have that

$$\|\Sigma_{S,\lambda}^{1/2}\widehat{\Sigma}_{S,\lambda}^{-1}\Sigma_{S,\lambda}^{1/2}\|_2 \le \frac{1}{1-\|\Delta_{S,\lambda}\|_2}.$$

□

Now, fix a subset S, we will drop the subscript for clearer presentation. For $z \in \mathbb{R}^{2s}$,

$$\begin{aligned} d_{1,\lambda} &:= \sum_{j=1}^{2s} \frac{\lambda_i(\Sigma_{SS})}{\lambda + \lambda_i(\Sigma_{SS})} \leq \sum_{j=1}^{2s} \frac{\lambda_i(\Sigma_{SS})}{\lambda} \leq \frac{2s}{\lambda} \\ \rho_\lambda &\geq \frac{\|\Sigma_{SS,\lambda}^{-1/2} z\|}{\sqrt{d_{1,\lambda}}} \\ \widetilde{d}_{1,\lambda} &:= \max\{1, d_{1,\lambda}\}. \end{aligned} \tag{32}$$

Note that, as $\|z\|_\infty \leq 1$, it means $\|z\|_2 \leq \sqrt{2s}$ we can choose $\rho_\lambda = \frac{\sqrt{2s}}{\sqrt{\lambda d_{1,\lambda}}}$.

Lemma C.4 ([Hsu et al., 2014]'s Lemma 2)**.** *For any δ such that $\ln(1/\delta) > \max\{0, 6 - \log \widetilde{d}_{1,\lambda}\}$, then with probability at least $1-\delta$, we have that*

$$\|\Delta_\lambda\|_2 \leq \sqrt{\frac{4\rho_\lambda^2 d_{1,\lambda}(\log(\widetilde{d}_{1,\lambda} + \delta^{-1}))}{N}} + \frac{2\rho_\lambda^2 d_{1,\lambda}(\log(\widetilde{d}_{1,\lambda} + \delta^{-1}))}{3n}. \tag{33}$$

With the choice of ρ_λ above, the bound can be further simplified as

$$\|\Delta_\lambda\|_2 \leq \sqrt{\frac{8s(\log(\widetilde{d}_{1,\lambda} + \delta^{-1}))}{\lambda n}} + \frac{4s(\log(\widetilde{d}_{1,\lambda} + \delta^{-1}))}{3\lambda n}. \tag{34}$$

Choose $\lambda = \frac{\sigma\sqrt{s}}{\sqrt{N}\|w_\star\|_2}$ to optimise the bound in Lemma C.1. Moreover, Taking union bound for all S, for $\|\Delta_\lambda\|_2 \leq 1/2$ with probability at least $1-\delta$, it suffices to choose $N = \Omega\left(\frac{s\|w_\star\|_2^2(\log(2s + d\delta^{-1}))}{\sigma^2(1-2\epsilon)^2}\right)$.

Proposition 3.2. *For any covariate matrix Σ, with probability at least $1-\delta$,there exists an estimator named* `SRLE2` *returns $\widehat{w}$ such that*

$$\|\widehat{w} - w_\star\|_\Sigma^2 = O\left(\frac{\sigma\|w_\star\|_2\sqrt{s}\log(d/\delta)}{\sqrt{N}} + \sigma^2\epsilon\right). \tag{4}$$

Proof. By choosing $\lambda = \frac{\sigma\sqrt{s}}{\sqrt{N}\|w_\star\|_2}$ and ensuring that $N = \Omega\left(\frac{s\|w_\star\|_2^2\log^2(2s + d\delta^{-1})}{(1-2\epsilon)^2}\right)$, we can guarantee that for all $S \subseteq [d]$ with $|S| \leq 2s$, $\Delta_{S,\lambda} \geq \frac{1}{2}$ with probability at least $1-\delta$. Combining the result in Lemma C.3 and Lemma C.1, we conclude the proof. □

C.3 Estimation Error of `SRLE3`

Similar as `SRLE1`, `SRLE3` was proposed in [Merad and Gaïffas, 2022] to solve the convex optimization problem

$$\begin{aligned} &\min_{w \in \mathbb{R}^d} \|Xw - Y\|_2^2 \\ &\text{s.t. } \|w\|_1 \leq \|w_\star\|_1, \end{aligned} \tag{35}$$

Without the uniform coverage assumption, the mirror descent algorithm with a suitable choice of Bregman divergence function is still able to achieve the slow rate. For further details, we refer the reader to the original paper [Merad and Gaïffas, 2022]. We now state the error guarantee of this estimator in our setting.

Proposition 3.3. *For any covariate matrix Σ, there exists an algorithm named* `SRLE3` *such that, with probability*

at least $1-\delta$*, it returns an estimate* $\widehat{w}$ *satisfying*

$$\|\widehat{w}-w_\star\|_\Sigma^2 = O\Bigg(\|w_\star\|_1(\|w_\star\|_1+\sigma) \times\left(\sqrt{\frac{\log(d\delta^{-1})}{N}}+\sqrt{\epsilon}\right)\Bigg).$$

Moreover, the estimator runs in $\mathrm{poly}(N,d,s)$ *time.*

Proof. By Proposition 3 in [Merad and Gaïffas, 2022], we have that

$$\|\widehat{w}-w_\star\|_\Sigma^2 \le \frac{\nu RL}{T}+4\|w_\star\|\sigma_{\max}\left(\sqrt{\frac{\log(d)+\log(\delta^{-1})}{N}}+\sqrt{\epsilon}\right). \tag{36}$$

Where $\nu=\frac{1}{2}e^2\log(d)$; and $L=1$ is the Lipschitz-smoothness, that is,

$$\|\Sigma(w-w')\|_\infty = \max_{i\in[d]}\langle\Sigma_i, w-w'\rangle \le \max_{i\in[d]}\|\Sigma_i\|_\infty\|w-w'\|_1 \le L\|w-w'\|_1.$$

and $\nu=\frac{1}{2}e^2\log(d)$. Moreover,

$$\sigma_{\max}^2 \triangleq \max_{w:\|w\|_1\le\|w_\star\|_1}\max_{j\in[d]}\mathbb{E}_{z,y}\left[((z^\top w-y)x_j)^2\right].$$

We note that, since $\|z\|_\infty\le 1$

$$\sigma_{\max}^2 \le \max_w \mathbb{E}_{z,y}\left[(z^\top w-y)^2\right] \le \max_w\max_z \mathbb{E}_{z,y}\left[(z^\top(w-w_\star))^2\right]+\mathbb{E}[\eta^2] = 4\|w_\star\|_1^2+\sigma^2.$$

□

D Proof of Section 4

D.1 Proof of Section 4.1

Proposition 4.1. *Suppose Assumption 2.1 and 2.2 hold. Let the pessimistic bonus be* $\Gamma_h(x,a)=0$ *for all* $h\in[H]$*,* $x\in\mathcal{X}$*,* $a\in\mathcal{A}$*. Run Algorithm 1 with the* `SRLE1` *estimator. Then, with probability at least* $1-\delta$*.*

$$\mathrm{SubOpt}(\pi_\star,\widehat{\pi}) = O\left(\frac{H^3 s\log(d/\delta)}{\xi\sqrt{N}}+\frac{H^3 s\sqrt{\epsilon}}{\xi}\right). \tag{7}$$

Proof. First, by Claim B.1, for any $Q_{h+1}\in\mathcal{Q}_{h+1}$, $\mathcal{P}_h^\pi(Q_{h+1})$ is s-sparse, and $\|\mathcal{P}_h^\pi(Q_{h+1})\|_1\le(H-h+1)$. Moreover, by Claim B.2, for any trajectory τ, we have $|R_h+\max_{a'\in\mathcal{A}}Q_{h+1}(x',a')|\le H$. Note that $|\mathcal{D}_h|=N/H$ and the worst-case contamination level in $\mathcal{D}_h$ is $H\epsilon$. Therefore, using the estimation error in Proposition 3.1 with $\|w\|_1\le 2Hs$ and $\sigma\le H$, we obtain

$$\|\mathcal{R}_h^*(\underline{Q}_{h+1})-\mathcal{P}_h^*(\underline{Q}_{h+1})\|_1 = O\left(\frac{H^{\frac{3}{2}}s\log(d/\delta)}{\xi\sqrt{N}}+\frac{H^2 s\sqrt{\epsilon}}{\xi}\right). \tag{37}$$

Similar to Lemma 3.1 of Zhang et al. [2021], we have that

$$\begin{aligned}\mathrm{SubOpt}(\pi,\widehat{\pi}) &\le 2H\max_{(x,a,h)}\left|\underline{Q}_h(x,a)-(\mathbb{B}_h\underline{Q}_{h+1})(x,a)\right| \\ &\le 2H\|\phi(x,a)\|_\infty\left\|\mathcal{R}_h^*(\underline{Q}_{h+1})-\mathcal{P}_h^*(\underline{Q}_{h+1})\right\|_1 \\ &= O\left(\frac{H^3 s\log(d/\delta)}{\xi\sqrt{N}}+\frac{H^3 s\sqrt{\epsilon}}{\xi}\right).\end{aligned}$$

□

D.2 Proof of Section 4.2

Proposition 4.4. *Let $\Sigma = \lambda I$. Let $z \sim \mathrm{Ber}^d(1/2)$. Then, for $d > 4s$, we have*

$$\underbrace{\mathbb{E}_{z\sim\mu}\Big[\max_{S:|S|=2s} z_S^\top \Sigma^{-1} z_S\Big]}_{LHS\ in\ (9)} - \underbrace{\max_{S:|S|=2s} \mathbb{E}_{x\sim\mu}\big[z_S^\top \Sigma^{-1} z_S\big]}_{RHS\ in\ (9)} \geq (1 - 2\exp(-d/8))\frac{s}{\lambda}. \tag{10}$$

Proof. For any subset $S \subset [d]$ with $|S| = 2s$ define

$$Z_S := \sum_{i\in S} z_i \qquad (\text{so } Z_S = z_S^\top z_S).$$

Step 1. Because each $z_i \in \{0,1\}$, Z_S counts how many of the $2s$ chosen coordinates are equal to 1. For a realisation of z let $K = \sum_{i=1}^d z_i$ (the *total* number of ones). If $K \geq 2s$ we can pick all $2s$ ones, so $\max_{|S|=2s} Z_S = 2s$. If $K < 2s$ we pick all K ones and fill the remaining slots with zeros, so the maximum equals K. Thus,

$$\max_{|S|=2s} Z_S = \min\{2s, K\}.$$

Step 2. Write $\mu = d/2$ for the mean of K. Because $2s \leq d/2 = \mu$, Chernoff's inequality gives

$$\mathbb{P}(K < 2s) \;=\; \mathbb{P}\big(K < (1-\delta)\mu\big) \;\leq\; \exp(-\tfrac{\delta^2\mu}{2}), \quad \delta := 1 - \tfrac{4s}{d} \;\in (0,1].$$

In particular, for all $d \geq 4s$ we have $\delta \geq \frac{1}{2}$, hence

$$\mathbb{P}(K < 2s) \;\leq\; e^{-d/8}.$$

Step 3. Using (1) we decompose

$$\mathbb{E}\Big[\max_{|S|=2s} Z_S\Big] = 2s\,\mathbb{P}(K \geq 2s) \;+\; \mathbb{E}\big[K\,\mathbf{1}_{\{K<2s\}}\big] \geq 2s(1 - e^{-d/8}).$$

If we use the ridge matrix $\Sigma = \lambda I_d$ (so that $\Sigma^{-1} = \frac{1}{\lambda} I_d$), then $z_S = \frac{1}{\lambda} Z_S$. Multiply inequality (2) by λ:

$$\mathbb{E}\Big[\max_{|S|=2s} z_S\Big] \;-\; \max_{|S|=2s} \mathbb{E}[z_S] \geq \lambda s\big(1 - 2e^{-d/8}\big); \qquad (d \geq 4s)$$

□

E Proof of Section 5

Induced MDP. To prove the results stated in Section 5, we use the notion of an induced MDP with perturbed rewards [Zanette et al., 2021b]. This perspective allows us to interpret the critic's pessimistic estimates as defining a modified MDP, in which the value functions coincide exactly with the pessimistic predictions.

Define the induced MDP $\widehat{M}(\pi) = (\mathcal{X}, \mathcal{A}, P, \widehat{r}^\pi, H)$ associated with $\{\underline{w}_h\}_{h=1}^H$, which differs from the original MDP M only in its reward function. For any (x,a), define

$$\widehat{r}_h^\pi(x,a) \triangleq r_h(x,a) + \underline{Q}_h^\pi(x,a) - (\mathbb{B}_h^\pi \underline{Q}_{h+1}^\pi)(x,a). \tag{38}$$

This construction is crucial: the linearity of the π-Bellman operator guarantees that the induced MDP is well-defined and enables tight pessimistic evaluation without the need for pointwise bonuses.

Below, we demonstrate that the induced MDP corresponds to the pessimistic evaluation of Q-function provided by the critic.

Proposition E.1. *$\widehat{M}(\pi)$ satisfies the following: $Q_{h,\widehat{M}^\pi}(x,a) = \underline{Q}_h^\pi(x,a)$, and $V_{h,\widehat{M}^\pi}(x) = \underline{V}_h^\pi(x)$.*

Proof. We begin by bounding the pessimistic property of the value function at the initial state:

$$Q^\pi_{h,\widehat{M}^\pi}(x,a) - Q^\pi_h(x,a) = \sum_{l=h}^{H} \mathbb{E}_{d_l^\pi}[\widehat{r}^\pi(x_l,a_l) - r(x_l,a_l)]. \tag{39}$$

On the other hand, using the definition of $\underline{Q}_h^\pi$ and the Bellman operator, we obtain

$$\begin{aligned}
\underline{Q}^\pi_h(x,a) - Q^\pi_h(x,a) &= \langle \phi(x,a), \underline{w}_h^\pi \rangle - \mathbb{B}_h^\pi(Q^\pi_{h+1})(x,a) \\
&= \Big(\langle \phi(x,a), \underline{w}_h^\pi \rangle - \mathbb{B}_h^\pi(\underline{Q}^\pi_{h+1})(x,a)\Big) + \Big(\mathbb{B}_h^\pi(\underline{Q}^\pi_{h+1})(x,a) - \mathbb{B}_h^\pi(Q^\pi_{h+1})(x,a)\Big) \\
&= \widehat{r}_h^\pi(x,a) - r_h(x,a) + \mathbb{E}_{x'\sim P(\cdot|x,a)}\mathbb{E}_{a\sim\pi(\cdot|x)}[\underline{Q}^\pi_{h+1} - Q^\pi_{h+1}](x,a).
\end{aligned} \tag{40}$$

Here, the last equality follows from the linearity of the Bellman operator $\mathbb{B}_h^\pi$.

Applying this argument recursively for $l = 1, \ldots, H$, we obtain

$$\underline{Q}^\pi_h(x,a) - Q^\pi_h(x,a) = \sum_{l=h}^{H} \mathbb{E}_{d_l^\pi}[\widehat{r}_l^\pi(x_l,a_l) - r_l(x_l,a_l)]. \tag{41}$$

This establishes the claim and concludes the proof. □

Next, we recall the general theorem for actor convergence [Zanette et al., 2021b], which will be used in our proof.

Theorem E.2 ([Zanette et al., 2021b])**.** *Let $M_t \triangleq \widehat{M}(\pi_t)$. For each M_t, define the advantage function*

$$G^{\pi_t}_{h,M_t}(x,a) \triangleq Q^{\pi_t}_{h,M_t}(x,a) - V^{\pi_t}_{h,M_t}(x).$$

Assume that the advantage function is uniformly bounded, i.e.,

$$\max_{t\in[T]} \max_{(x,a,h)} |G^{\pi_t}_{h,M_t}(x,a)| \le B.$$

Suppose further that $T \ge \log(|\mathcal{A}|)$ and that the step size satisfies $\eta \in (0,1)$. Then, for any fixed policy π, we have

$$\frac{1}{T}\sum_{t=1}^{T}\Big(V^{\pi_t}_{1,M_t}(x_1) - V^{\pi}_{1,M_t}(x_1)\Big) = O\left(H\left(\frac{\log(|\mathcal{A}|)}{\eta T} + \eta B^2\right)\right). \tag{42}$$

E.1 Proof of Section 5.1

For a sequence $\beta = (\beta_h)_{h=1}^H$ and any given policy π, define the good event

$$\mathcal{G}'^\pi(\beta) \triangleq \left\{ \sup_{Q_{h+1}\in\mathcal{Q}_{h+1}} \|\mathcal{E}_h^\pi(Q_{h+1})\|_1 \le \beta_h,\ \forall h \in [H] \right\}. \tag{43}$$

Under the good event $\mathcal{G}'^\pi(\beta)$ for the sequence of policies produced by the actor, we will show that the suboptimality gap is small. Before doing so, we first establish that, under the good event $\mathcal{G}'^\pi(\beta)$, the critic's estimation error can be bounded as follows.

Proposition E.3. *Conditioned on the event $\mathcal{G}'^\pi(\beta)$, for some input policy π, the critic returns an induced MDP*

$\widehat{M}(\pi)$ *such that, for every policy* $\widetilde{\pi}$,

$$\left|V^{\widetilde{\pi}}_{1,\widehat{M}(\pi)}(x_1) - V^{\widetilde{\pi}}_1(x_1)\right| \leq 2\sum_{h=1}^{H}\beta_h. \tag{44}$$

Proof. Let $\widehat{w}^\pi_h = \mathcal{R}^\pi(\underline{Q}^\pi_{h+1})$. We bound $\underline{w}^\pi_h - \mathcal{P}^\pi_h(\underline{Q}^\pi_{h+1})$ as follows:

$$\begin{aligned}
\left\|\underline{w}^\pi_h - \mathcal{P}^\pi_h(\underline{Q}^\pi_{h+1})\right\|_1 &\leq \|\underline{w}^\pi_h - \widehat{w}^\pi_h\|_1 + \left\|\widehat{w}^\pi_h - \mathcal{P}^\pi_h(\underline{Q}^\pi_{h+1})\right\|_1 \\
&\leq 2\left\|\widehat{w}^\pi_h - \mathcal{P}^\pi_h(\underline{Q}^\pi_{h+1})\right\|_1 \qquad \text{(Conditioned on } \mathcal{G}'^\pi(\beta),\ \widehat{w}^\pi_h \text{ is a feasible solution of (12))} \\
&\leq 2\beta_h.
\end{aligned}$$

Therefore, for all $h \in [H]$,

$$\begin{aligned}
\max_{(x,a)} \widehat{r}^\pi_h(x,a) - r_h(x,a) &= \max_{(x,a)} \left|\underline{Q}^\pi_h(x,a) - (\mathbb{B}^\pi \underline{Q}^\pi_{h+1})(x,a)\right| \\
&\leq \left\langle \underline{w}^\pi_h - \mathcal{P}^\pi(\underline{Q}^\pi_h), \phi(x,a)\right\rangle \\
&\leq \left\|\underline{w}^\pi_h - \mathcal{P}^\pi(\underline{Q}^\pi_h)\right\|_1 \|\phi(x,a)\|_\infty \\
&\leq 2\beta_h.
\end{aligned}$$

Therefore, we have that

$$V^{\widetilde{\pi}}_{1,\widehat{M}(\pi)}(x_1) - V^{\widetilde{\pi}}_1(x_1) = \sum_{h=1}^{H}\mathbb{E}_{d^{\widetilde{\pi}}}[\widehat{r}^\pi_h(x,a) - r_h(x,a)] \leq \sum_{h=1}^{H}\left(\max_{(x,a)}\widehat{r}^\pi_h(x,a) - r_h(x,a)\right) \leq 2\sum_{h=1}^{H}\beta_h.$$

This completes the proof. □

Theorem 5.1. *Suppose Assumption 2.1 and 2.2 hold. Then, with step size* $\eta = \sqrt{\frac{\log \mathcal{A}}{N}}$, *after* $T = N/H$ *iterations the actor-critic algorithm returns a policy* $\widehat{\pi}$ *that satisfies*

$$\begin{aligned}
\mathrm{SubOpt}(\pi_\star, \widehat{\pi}) = O\Bigg(&\frac{H^3 s \log(dNH\delta^{-1})}{\xi\sqrt{N}} \\
&+ H^3\sqrt{\frac{\log(|\mathcal{A}|)}{N}} + \frac{H^3 s\sqrt{\epsilon}}{\xi}\Bigg),
\end{aligned} \tag{13}$$

with probability at least $1-\delta$.

Proof. **Bound probability of good event** $\mathcal{G}'(\beta)$**.** choose the sequence β

$$\beta_h = O\left(\frac{H^2 s\log(dHN\delta^{-1})}{\xi\sqrt{N}} + \frac{H^2 s\sqrt{\epsilon}}{\xi}\right).$$

According to Proposition 3.1, and by Claim B.1 and Claim B.2, we have that for any policy π, $\mathbb{P}(\mathcal{G}'^\pi(\beta)) \geq 1-\delta/T$. Therefore, for any sequence of policy $(\pi_t)_{t=1}^T$, taking union bound for T policies and H horizon, we have that, $\mathbb{P}(\cap_{t=1}^T \mathcal{G}'^{\pi_t}(\beta)) \geq 1-\delta$.

Suboptimality gap decomposition. Let $M_t \triangleq \widehat{M}(\pi_t)$. From the critic analysis, under event $\mathcal{G}^{\pi_t}(\beta)$,

$$\begin{aligned} V_1^{\pi_\star}(x_1) - V_1^{\pi_t}(x_1) &= \left(V_1^{\pi_\star}(x_1) - V_{1,M_t}^{\pi_\star}(x_1)\right) + \left(V_{1,M_t}^{\pi_\star}(x_1) - V_{1,M_t}^{\pi_t}(x_1)\right) + \left(V_{1,M_t}^{\pi_t}(x_1) - V_1^{\pi_t}\right) \\ &\le V_{1,M_t}^{\pi_\star}(x_1) - V_{1,M_t}^{\pi_t}(x_1) + 2\sum_{h=1}^{H}\beta_h. \end{aligned} \tag{45}$$

Actor Convergence. We now bound

$$\frac{1}{T}\sum_{t=1}^{T}\left[V_{1,M_t}^{\pi_\star}(x_1) - V_{1,M_t}^{\pi_t}(x_1)\right].$$

To do so, we invoke Theorem E.2 and provide an upper bound for the advantage function parameter B, namely $B = O(H)$.

By Proposition E.1, and assuming the event $\mathcal{G}^{\pi_t}(\beta)$ holds for the sequence $(\pi_t)_{t=1}^T$ with $\beta_h \le 1$, we obtain

$$|\widehat{r}_h^{\pi_t}(x,a) - \widehat{r}_h^{\pi_t}(x,a')| \le |r_h(x,a) - r_h(x,a')| + 2\beta_h \le 4.$$

Moreover, from the constraints in (12), for any (x,h) we have

$$V_{h+1,M_t}^{\pi_t}(x) \le H - h.$$

Therefore, for any triplet (x,a,h),

$$\begin{aligned} \left|Q_{h,M_t}^{\pi_t}(x,a) - Q_{h,M_t}^{\pi_t}(x,a')\right| &\le |\widehat{r}_h^{\pi_t}(x,a) - \widehat{r}_h^{\pi_t}(x,a')| + \left|\int_{x'\in\mathcal{X}} V_{h+1,M_t}^{\pi_t}(x')\big(P(x' \mid x,a) - P(x' \mid x,a')\big)\right| \\ &\le O(1) + \left|\left\langle \phi(x,a) - \phi(x,a'), \int_{x'\in\mathcal{X}} V_{h+1,M_t}(x')\mu(x')\right\rangle\right| \\ &\le O(1) + \|\phi(x,a) - \phi(x,a')\|_\infty \left\|\int_{x'\in\mathcal{X}} V_{h+1,M_t}(x')\mu(x')\right\|_1 \\ &\le O(1) + 2\left\|\int_{x'\in\mathcal{X}} V_{h+1,M_t}(x')\mu(x')\right\|_1 \\ &\le O(1) + 2\left\|H\int_{x'\in\mathcal{X}} \mu(x')\right\|_1 \\ &\le O(1) + 2H. \end{aligned}$$

Thus, we can take $B = O(H)$.

Applying Theorem E.2, we conclude that

$$\frac{1}{T}\sum_{t=1}^{T}\left[V_{1,M_t}^{\pi_\star}(x_1) - V_{1,M_t}^{\pi_t}(x_1)\right] \le 4H^2\sqrt{\frac{\log(|\mathcal{A}|)}{T}}. \tag{46}$$

Putting things together. Let $\widehat{\pi}$ be the mixture of policies $\{\pi_1, ..., \pi_T\}$ and $T = N/H$, we have

$$V_1^{\pi_\star}(x_1) - V_1^{\pi_t}(x_1) \le 2\sum_{h=1}^{H}\beta_h + 4H^3\sqrt{\frac{\log \mathcal{A}}{N}} \tag{47}$$

with probability at least $1-\delta$. Finally, by the choice of β, we have that

$$V_1^{\pi_\star}(x_1) - V_1^{\pi_t}(x_1) = O\left(\frac{H^3 s\log(dNH\delta^{-1})}{\xi\sqrt{N}} + H^3\sqrt{\frac{\log(|\mathcal{A}|)}{N}} + \frac{H^3 s\sqrt{\epsilon}}{\xi}\right), \tag{48}$$

with probability at least $1-\delta$. □

E.2 Proof of Section 5.2

First, we recall the definition of the good event used in this section, and show that it holds with high probability. For a sequence $\alpha = (\alpha_h)_{h=1}^H$ and any policy π, define

$$\mathcal{G}^\pi(\alpha) \triangleq \left\{ \sup_{Q_{h+1}\in\mathcal{Q}_{h+1}} \|\mathcal{E}_h^\pi(Q_{h+1})\|_{\widehat{\Sigma}_h}^2 \le \alpha_h, \ \ \forall h \in [H] \right\}. \tag{49}$$

Proposition E.4. *Run Algorithm 2 with the* `SRLE2` *and* `SRLE3` *estimators. Choose the sequences α and α' for these two estimators as follows:*

$$\alpha_h^2 = O\left(\frac{\sqrt{s}H^3 \log(dH\delta^{-1})}{\sqrt{N}} + H^3\epsilon + H^3(\lambda+\epsilon) \right),$$

and

$$\alpha_h'^2 = O\left(\frac{H^3 \log(dH\delta^{-1})}{\sqrt{N}} + H^3\sqrt{\epsilon} + H^3(\lambda+\epsilon) \right).$$

Then, for any policy π,

$$\mathbb{P}(\mathcal{G}^\pi(\alpha)) \ge 1-\delta,$$

when using the `SRLE2` *estimator, and*

$$\mathbb{P}(\mathcal{G}^\pi(\alpha')) \ge 1-\delta,$$

when using the `SRLE3` *estimator.*

Proof. We first prove the result for the `SRLE2` estimator. The proof for the `SRLE3` estimator follows the same argument, with Proposition 3.2 replaced by Proposition 3.3.

We apply the property of the sparse linear regression oracle. By Claim B.1, for any policy π,

$$\|\mathcal{P}_h^\pi(Q_{h+1})\|_1 \le H-h, \qquad \|\mathcal{P}_h^\pi(Q_{h+1})\|_0 \le s. \tag{50}$$

Let $z_h^\tau = \phi(x_\tau, a_\tau)$ and define

$$y_h^\tau = r_h^\tau + \sum_{a\in\mathcal{A}} \pi(a \mid x_{h+1}^\tau) Q_{h+1}(x_{h+1}^\tau, a).$$

Then, by Claim B.2,

$$\mathbb{E}[y_h^\tau] = \langle \phi(x_\tau, a_\tau), \mathcal{P}_h^\pi(Q_{h+1}) \rangle, \qquad |y_h^\tau| \le H-h.$$

Note that dual to data spliting, the worst case contamination level in set $\mathcal{D}_h$ is $H\epsilon$. By Proposition 3.2, for all $h\in[H]$, the `SRLE2` oracle ensures that

$$\|\mathcal{E}_h^\pi(Q_{h+1})\|_{\Sigma_h}^2 = O\left(\frac{H^3\sqrt{s}\log(dH/\delta)}{\sqrt{N}} + H^3\epsilon \right),$$

with probability at least $1-\delta$.

Claim E.5.

$$\|\mathcal{E}_h^\pi(Q_{h+1})\|_{\widehat{\Sigma}_h}^2 = O\big(\|\mathcal{E}_h^\pi(Q_{h+1})\|_{\Sigma_h}^2 + H^3(\epsilon+\lambda) \big). \tag{51}$$

Proof of claim. Let $e = \mathcal{E}_h^\pi(Q_{h+1})$. Note that, due to data splitting, Q_{h+1} does not dependent on data set $\mathcal{D}_h$.

We have $\|e\|_1 = O(H-h)$ and $\|e\|_0 \le 2s$. Let S denote the sparsity support of e. Then

$$\begin{aligned}\|e\|^2_{\widehat{\Sigma}_h} &= e^\top \widehat{\Sigma}_h e \\ &= e_S^\top \left[\frac{H}{N}\sum_{i=1}^{N/H}\phi_i\phi_i^\top + (\lambda+\epsilon)I\right] e_S \\ &= e_S^\top \left[\underbrace{\frac{H}{N}\sum_{i=1}^{(1-\epsilon)N/H}\widetilde{\phi}_i\widetilde{\phi}_i^\top}_{(1-\epsilon)\widetilde{\Sigma}_h,\text{ clean data}} + \underbrace{\frac{H}{N}\sum_{i=1}^{\epsilon N/H}\phi_i'\phi_i'^\top}_{\text{corrupted data}} + (\lambda+\epsilon)I\right] e_S.\end{aligned} \tag{52}$$

For the corrupted data,

$$\frac{H}{N}\sum_{i=1}^{\epsilon N/H} e^\top\phi_i'\phi_i'^\top e \le \frac{H}{N}\sum_{i=1}^{\epsilon N/H}(\|\phi_i'\|_\infty\|e\|_1)^2 \le \epsilon\|e\|_1^2 \le O((H-h)^2\epsilon).$$

For the clean data, by Lemma F.1, for $0<\epsilon<1/2$,

$$\left[\underbrace{\frac{H}{N}\sum_{i=1}^{(1-\epsilon)N/H}\widetilde{\phi}_i\widetilde{\phi}_i^\top}_{(1-\epsilon)\widetilde{\Sigma}_h,\text{ clean data}} + (\lambda+\epsilon)I\right]_S = O([\Sigma_h + (\lambda+H\epsilon)I]_S).$$

Therefore,

$$\|e\|^2_{\widehat{\Sigma}_h} = O(\|e\|^2_{\Sigma_h} + H^3(\lambda+\epsilon)).$$

□

Combining the results completes the proof of the proposition. □

Proposition E.6 (Policy evaluation)**.** *Condition on the event $\mathcal{G}^\pi(\alpha)$, when given any policy π, the critic returns an induced MDP $\widehat{M}(\pi)$ such that:*
[a] *For the given policy π, we have*

$$V^\pi_{1,\widehat{M}(\pi)}(x_1) \le V_1^\pi(x_1). \tag{53}$$

[b] *For any policy $\widetilde{\pi}$, we have*

$$\left|V^{\widetilde{\pi}}_{1,\widehat{M}(\pi)}(x_1) - V_1^{\widetilde{\pi}}(x_1)\right| \le 2\sum_{h=1}^H \alpha_h \mathbb{E}_{d^{\widetilde{\pi}}}\left[\left\|[\phi(x,a)]_{\widetilde{S}_h}\right\|_{\widehat{\Sigma}_h^{-1}}\right]. \tag{54}$$

Proof. **Part a:** We first show that the ground truth $\widetilde{w}_h^\pi$, that is, the ground truth $Q_h^\pi(x,a) = \langle\phi(x,a), \widetilde{w}_h^\pi\rangle$, satisfies the program. In particular, let $\widetilde{w}_{H+1}^\pi = \mathbf{0}$ and $\widetilde{w}_h^\pi = \mathcal{P}_h^\pi(\langle\phi, \widetilde{w}_{h+1}^\pi\rangle)$. By Claim B.1, we have that $\|\widetilde{w}_h^\pi\|_1 \le H-h$.

Next, under the event $\mathcal{G}^\pi(\alpha)$, and note that ground truth $Q_{h+1}^\pi \in \mathcal{Q}_{h+1}$, we have that

$$\|\widetilde{w}_h^\pi - \mathcal{R}_h^\pi(Q_{h+1}^\pi)\|^2_{\widehat{\Sigma}_h} \le \alpha_h.$$

Which means that $(\widetilde{w}_h^\pi)_{h=1}^H$ is a feasible solution of the optimization problem. The output of the optimization problem $\underline{w}_h^\pi$ satisfies

$$V^\pi_{1,M(\pi)}(x_1) = \underline{V}_1^\pi(x_1) \le V_1^\pi(x_1). \tag{55}$$

Part b: First, from the definition of perturbed reward of $M(\pi)$, we have

$$\begin{aligned}\widehat{r}_h^\pi(x,a) - r_h(x,a) &= \langle \phi_h(x,a), \underline{w}_h^\pi \rangle - \mathbb{B}_h^\pi(\underline{Q}_{h+1}^\pi)(x,a) \\ &= \langle \phi_h(x,a), \underline{w}_h^\pi - \mathcal{R}(\underline{Q}_{h+1}^\pi)\rangle + \mathcal{R}(\underline{Q}_{h+1}^\pi - \mathbb{B}_h^\pi(\underline{Q}_{h+1}^\pi)(x,a) \\ &= \langle \phi_h(x,a), \underline{w}_h^\pi - \mathcal{R}(\underline{Q}_{h+1}^\pi)\rangle + \langle \phi_h(x,a), \mathcal{R}(\underline{Q}_{h+1}^\pi) - \mathcal{P}(\underline{Q}_{h+1}^\pi)\rangle \\ &\leq 2\|[\phi(x,a)]_{\widetilde{S}}\|_{\widehat{\Sigma}_h^{-1}}\alpha_h.\end{aligned} \tag{56}$$

Therefore,

$$\begin{aligned}\left|V_{1,\widehat{M}(\pi)}^{\widetilde{\pi}}(x_1) - V_1^{\widetilde{\pi}}(x_1)\right| &= \left|\sum_{h=1}^H \mathbb{E}_{d_h^{\widetilde{\pi}}}[\widehat{r}_h^\pi(x,a) - r_h(x,a)]\right| \\ &\leq \left|2\sum_{h=1}^H \mathbb{E}_{d_h^{\widetilde{\pi}}}\left[\alpha_h \left\|[\phi(x,a)]_{\widetilde{S}}\right\|_{\widehat{\Sigma}_h^{-1}}\right]\right|.\end{aligned} \tag{57}$$

□

Proposition E.7 (Actor's convergence)**.** *For the sequence $(\pi_t)_{t=1}^T$, suppose that the event $\mathcal{G}^{\pi_t}(\alpha)$ holds for all $t \in [T]$. Suppose the actor takes $T \geq \log|\mathcal{A}|$ steps with step size $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{TH^2}}$. Then, for any policy π,*

$$\frac{1}{T}\sum_{t=1}^T\left(V_{1,M_t}^\pi(x_1) - V_{1,M_t}^{\pi_t}(x_1)\right) = O\left(H^2\sqrt{\frac{\log(|\mathcal{A}|)}{T}}\right).$$

Proof. The proof is Proposition E.7 followed by the general result for offline actor-critic as stated in Theorem E.2. The Proposition E.7 is immediately followed if we can show the constant in the theorem $B = O(H)$ and choosing the stepsize η. First, we bound the magnitude of the advantage function $G_{h,M_t}^{\pi_t}$. Note that by the definition of induced MDP, M_t is only different from the original MDP in the reward function. By Proposition E.1, and suppose the event $\mathcal{G}^{\pi_t}(\alpha)$ holds for given sequence $(\pi_t)_{t=1}^T$ with $\alpha_h \leq 1$, we have that

$$|\widehat{r}_h^{\pi_t}(x,a) - \widehat{r}_h^{\pi_t}(x,a')| \leq |r_h(x,a) - r_h(x,a')| + 2\alpha_h \leq 4.$$

Moreover, thanks to the pessimistic property of induced MDP in Proposition E.6, for any (x,h), $V_{h,M_t}^{\pi_t}(x) \leq V_h^{\pi_t}(x) \leq H - h + 1$. Therefore, for any triplet (x,a,h), we have that

$$\begin{aligned}\left|Q_{h,M_t}^{\pi_t}(x,a) - Q_{h,M_t}^{\pi_t}(x,a')\right| &\leq |\widehat{r}_h^{\pi_t}(x,a) - \widehat{r}_h^{\pi_t}(x,a')| + \left|\int_{x'\in\mathcal{X}} V_{h+1,M_t}^{\pi_t}(x')(P(x' \mid (x,a) - P(x' \mid (x,a'))\right| \\ &\leq O(1) + \left|\left\langle \phi(x,a) - \phi(x,a'), \int_{x'\in\mathcal{X}} V_{h+1,M_t}(x')\mu(x')\right\rangle\right| \\ &\leq O(1) + \|\phi(x,a) - \phi(x,a')\|_\infty \left\|\int_{x'\in\mathcal{X}} V_{h+1,M_t}(x')\mu(x')\right\|_1 \\ &\leq O(1) + 2\left\|\int_{x'\in\mathcal{X}} V_{h+1,M_t}(x')\mu(x')\right\|_1 \\ &\leq O(1) + 2\left\|H\int_{x'\in\mathcal{X}} \mu(x')\right\|_1 \\ &\leq O(1) + 2H.\end{aligned}$$

Therefore, we can choose $B = O(H)$. Now, choose $\eta = \frac{\sqrt{\log(|\mathcal{A})|}}{H\sqrt{T}}$, we obtain the result. □

Theorem 5.3. *Suppose Assumption 2.1 and 2.3 hold. Let operators $(\mathcal{R}_h^\pi)_{h=1}^H$ be defined using the* SRLE2 *estimator.*

Choose the sequence α such that

$$\alpha_h^2 = O\left(\frac{\sqrt{s}H^3\log(dH\delta^{-1})}{\sqrt{N}} + H^3\epsilon + H^3(\lambda+\epsilon)\right).$$

Choose $\lambda = \frac{s}{N}\log\frac{d}{s\delta}$. Then, after $T = N$ iterations with step size $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{H^2N}}$, the actor-critic algorithm returns a policy $\widehat{\pi}$ that satisfies

$$\begin{aligned}\mathrm{SubOpt}(\pi_\star, \widehat{\pi}) = O\Bigg(&\frac{H^3\sqrt{\kappa}s^{3/4}\sqrt{\log(dHN\delta^{-1})}}{N^{\frac{1}{4}}} \\ &+ H^3\sqrt{\frac{\log(|\mathcal{A}|)}{N}} + H^3\sqrt{\kappa s\epsilon}\Bigg),\end{aligned}$$

with probability at least $1-\delta$.

Proof. We note that, for any sequence of T policy $(\pi_t)_{t=1}^T$, with prescribed α, the event $\mathbb{P}(\cap_{t=1}^T \mathcal{G}^{\pi_t}(\alpha)) \geq 1-\delta$.

Next, from the critic analysis, under event $\cap_{t=1}^T \mathcal{G}^{\pi_t}(\alpha)$, we have that

$$\begin{aligned}V_1^{\pi_\star}(x_1) - V_1^{\pi_t}(x_1) &\leq \left(V_{1,M_t}^{\pi_\star}(x_1) + 2\sum_{h=1}^H \alpha_h \mathbb{E}_{d^{\pi_\star}}\left[\left\|[\phi(x,a)]_{\widetilde{S}_h}\right\|_{\widehat{\Sigma}_h^{-1}}\right]\right) - V_{1,M_t}^{\pi_t}(x_1) \\ &= V_{1,M_t}^{\pi_\star}(x_1) - V_{1,M_t}^{\pi_t}(x_1) + 2\sum_{h=1}^H \alpha_h \mathbb{E}_{d^{\pi_\star}}\left[\left\|\phi(x,a)]_{\widetilde{S}_h}\right\|_{\widehat{\Sigma}_h^{-1}}\right].\end{aligned} \tag{58}$$

As $\epsilon \leq 1/2$, by Lemma F.1, with number of samples $N = \Omega(s\log(d/\delta))$, then with probability at least $1-\delta$, for all $\widetilde{S} \subset [d]$ such that $|\widetilde{S}| \leq 2s$, we have that

$$[\widehat{\Sigma}_h]_{\widetilde{S}} \geq \frac{1}{3}([\Sigma_h + \lambda I]_{\widetilde{S}}).$$

Now, consider

$$\begin{aligned}\mathbb{E}_{d^{\pi_\star}}[\|[\phi(x,a)]_{\widetilde{S}_h}\|_{\widehat{\Sigma}_h^{-1}}] &\leq \mathbb{E}_{d^{\pi_\star}}[\|[\phi(x,a)]_{\widetilde{S}_h}\|_{3(\Sigma_h+\lambda I)^{-1}}] \\ &\leq \mathbb{E}_{d^{\pi_\star}}\left[\sqrt{3\phi_{\widetilde{S}}(\Sigma_h+\lambda I)^{-1}\phi_{\widetilde{S}}}\right] \\ &\leq \sqrt{\mathbb{E}_{d^{\pi_\star}}\left[3\phi_{\widetilde{S}}(\Sigma_h+\lambda I)^{-1}\phi_{\widetilde{S}}\right]} \\ &= \sqrt{3\mathrm{Tr}([\Sigma_*\widehat{\Sigma}_h^{-1}]_{\widetilde{S}})} \\ &\leq \sqrt{3\kappa\mathrm{Tr}([\Sigma_h\widehat{\Sigma}_h^{-1}]_{\widetilde{S}})} \\ &= \sqrt{3\kappa\mathrm{Tr}\left(\sum_{i\in\widetilde{S}}\frac{\lambda_i}{\lambda_i+\lambda}\right)^{-1}} \\ &\leq \sqrt{6\kappa s}.\end{aligned} \tag{59}$$

Now, from actor analysis, we have that

$$\frac{1}{T}\sum_{t=1}^T\left[V_{1,M_t}^{\pi_\star}(x_1) - V_{1,M_t}^{\pi_t}(x_1)\right] \leq 4H^2\sqrt{\frac{\log|\mathcal{A}|}{T}} \tag{60}$$

Therefore, let $\widehat{\pi}$ be the mixture of policies $\{\pi_1, ..., \pi_T\}$. Let $T = N/H$. We have that

$$V_1^{\pi_\star}(x_1) - V_1^{\widehat{\pi}}(x_1) \leq 2\sqrt{6\kappa} \sum_{h=1}^{H} \alpha_h + 4H^3 \sqrt{\frac{\log |\mathcal{A}|}{N}}, \tag{61}$$

with probability at least $1 - \delta$. □

Theorem 5.4. *Suppose Assumption 2.1 and 2.3 hold. Let operator $\mathcal{R}_h^\pi$ be defined using the* `SRLE3` *estimator. Choose the sequence α such that*

$$\alpha_h = O\left(\frac{H^3 \log(dH\delta^{-1})}{\sqrt{N}} + H^3\sqrt{\epsilon} + H^3(\lambda + \epsilon) \right).$$

Choose $\lambda = \frac{s}{N} \log \frac{d}{s\delta}$. After $T = N$ iterations with step size $\eta = \sqrt{\frac{\log(|\mathcal{A}|)}{H^2 N}}$, the actor-critic algorithm returns a policy $\widehat{\pi}$ that satisfies

$$\begin{aligned} \text{SubOpt}(\pi_\star, \widehat{\pi}) = O\Bigg(& \frac{H^3 \sqrt{\kappa s} \sqrt{\log(dHN\delta^{-1})}}{N^{\frac{1}{4}}} \\ & + H^3 \sqrt{\frac{\log(|\mathcal{A}|)}{N}} + H^3 \sqrt{\kappa s} \epsilon^{\frac{1}{4}} \Bigg), \end{aligned} \tag{16}$$

with probability at least $1 - \delta$.

Proof. The proof of Theorem 5.4 follows the same argument as the proof of Theorem 5.3, using the value of α specified in its statement. We therefore omit the details for brevity. □

F Technical Lemmas

Lemma F.1 (Regularised covariance concentration on sparse supports)**.** *Let $\{\phi_i\}_{i=1}^N \subset \mathbb{R}^d$ be i.i.d. random vectors satisfying*

$$\|\phi_i\|_\infty \leq 1 \quad a.s.$$

Define the population and empirical (ridge-regularised) covariance matrices

$$\Sigma = \mathbb{E}[\phi\, \phi^\top], \qquad \widehat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} \phi_i \phi_i^\top + \lambda I_d.$$

Fix a sparsity level $s \in \{1, \ldots, d\}$ and let $S \subset [d]$ with $|S| \leq s$. Denote by Σ_S (resp. $\widehat{\Sigma}_S$) the principal $s \times s$ sub-matrix of Σ (resp. $\widehat{\Sigma}$) indexed by S.

There is an absolute numerical constant $C > 0$ such that, for every $\delta \in (0, 1)$, if the ridge parameter satisfies

$$\lambda = C \frac{s}{N} \log \frac{d}{s\delta}. \tag{62}$$

then, with probability at least $1 - \delta$,

$$\forall S \subset [d],\ |S| \leq s : \qquad \frac{1}{3} \left[\Sigma + \lambda I_d\right]_S \preceq \widehat{\Sigma}_S \preceq \frac{5}{3} \left[\Sigma + \lambda I_d\right]_S. \tag{63}$$

Proof. We adapt the argument of Lemma 39 in Appendix I of [Zanette et al., 2021a]; the main changes are the use of the ℓ_∞ bound and the restriction to supports of size at most s.

Step 1. Fix $S \subset [d]$ with $|S| \leq s$ and a unit vector $x \in \mathbb{R}^d$ supported on S. Because each coordinate of ϕ_i lies in $[-1, 1]$,

$$\left|x^\top \phi_i\right| = \left|\sum_{j \in S} x_j \phi_{ij}\right| \leq \sum_{j \in S} |x_j| \leq \sqrt{s}\, \|x\|_2 = \sqrt{s},$$

so $(x^\top \phi_i)^2 \in [0, s]$. Write $Z_i(x) = (x^\top \phi_i)^2 - \mathbb{E}[(x^\top \phi)^2]$ and $\mu(x) = x^\top \Sigma x$. Then $\{Z_i(x)\}_{i=1}^N$ are mean-zero, independent, $|Z_i(x)| \le s$ and $\mathrm{Var}[Z_i(x)] \le s\,\mu(x)$. Bernstein's inequality gives, for any $t > 0$,

$$\mathbb{P}\Big(\Big|\tfrac{1}{N}\sum_{i=1}^N Z_i(x)\Big| > t\Big) \le 2\exp\Big(-\frac{Nt^2}{2\mu(x) + \frac{2}{3}st}\Big). \tag{64}$$

Step 2: Choosing t and the ridge λ. Set

$$t(x) \;=\; \tfrac{1}{3}(\mu(x) + \lambda).$$

Under (62),

$$\frac{Nt(x)^2}{2\mu(x) + \frac{2}{3}st(x)} \;\ge\; \frac{N(\mu(x)+\lambda)^2}{18\mu(x) + 4s(\mu(x)+\lambda)} \;\ge\; C_1\,\frac{N\lambda}{s},$$

for an absolute C_1. Taking $C \ge 6C_1$ in (62) forces the exponent in (64) to exceed $s\log\frac{ed}{s} + \log\frac{2}{\delta}$. Thus, for the fixed x,

$$\mathbb{P}\Big(\big|x^\top(\widehat{\Sigma} - \Sigma)x\big| > \tfrac{1}{3}(\mu(x)+\lambda)\Big) \;\le\; 2\exp\big(-s\log\tfrac{ed}{s} - \log\tfrac{2}{\delta}\big).$$

Step 3: Uniformity over all x and all supports. Let $\mathcal{N}_\varepsilon^{(s)}$ be an ε-net of the unit sphere in $\mathbb{R}^s$, lifted to $\mathbb{R}^d$ by zero padding outside S. With $\varepsilon = \frac{1}{6}$ and a union bound over $|\mathcal{N}_\varepsilon^{(s)}| \le (3/\varepsilon)^s$ choices of x and $\binom{d}{s} \le (ed/s)^s$ subsets S, the preceding probability remains below δ. Standard arguments then show that on the resulting event

$$\big|y^\top(\widehat{\Sigma} - \Sigma)y\big| \;\le\; \tfrac{2}{3}\,\big[y^\top\Sigma y + \lambda\big], \quad \forall y \in \mathbb{R}^d,\ \|y\|_2 = 1,\ \mathrm{supp}(y) \subseteq S,$$

simultaneously for every $|S| \le s$.

Step 4. Let

$$D \;:=\; \frac{1}{N}\sum_{i=1}^N \phi_i\phi_i^\top \;-\; \Sigma \;=\; \widehat{\Sigma} - \Sigma - \lambda I_d,$$

so that $\widehat{\Sigma} = \Sigma + \lambda I_d + D$. From the outcome of Step 3 we know that the event

$$\big|y^\top D y\big| \;\le\; \frac{2}{3}\big[y^\top\Sigma y + \lambda\big] \quad \forall\, y \in \mathbb{R}^d,\ \|y\|_2 = 1,\ \mathrm{supp}(y) \subseteq S \tag{65}$$

occurs with probability at least $1 - \delta$. Fix such a vector y and write $q := y^\top\Sigma y + \lambda > 0$. From (65) we have

$$-\frac{2}{3}\,q \;\le\; y^\top D y \;\le\; \frac{2}{3}\,q.$$

Adding $q = y^\top\Sigma y + \lambda$ to every term yields

$$\frac{1}{3}\,q \;\le\; y^\top\widehat{\Sigma}y \;\le\; \frac{5}{3}\,q,$$

that is

$$\frac{1}{3}\,y^\top(\Sigma + \lambda I_d)y \;\le\; y^\top\widehat{\Sigma}y \;\le\; \frac{5}{3}\,y^\top(\Sigma + \lambda I_d)y \qquad \forall\, y \in \mathbb{R}^d,\ \|y\|_2 = 1,\ \mathrm{supp}(y) \subseteq S.$$

Since the inequality above is valid for all unit vectors supported on S, it is equivalent to the matrix bound

$$\frac{1}{3}\,\big[\Sigma + \lambda I_d\big]_S \;\preceq\; \widehat{\Sigma}_S \;\preceq\; \frac{5}{3}\,\big[\Sigma + \lambda I_d\big]_S$$

which is exactly the claim of Lemma F.1 for the principal sub-matrix indexed by S. □