Model-free Low-rank Reinforcement Learning via Leveraged Entry-wise Matrix Estimation

Stefan Stojanovic
KTH, Stockholm, Sweden
stesto@ktj.se

Yassir Jedra
MIT, Cambridge, USA
jedra@mit.edu

Alexandre Proutiere
KTH, Digital Futures, Stockholm, Sweden
alepro@kth.se

Abstract

We consider the problem of learning an ε-optimal policy in controlled dynamical systems with low-rank latent structure. For this problem, we present `LoRa-PI` (Low-Rank Policy Iteration), a model-free learning algorithm alternating between policy improvement and policy evaluation steps. In the latter, the algorithm estimates the low-rank matrix corresponding to the (state, action) value function of the current policy using the following two-phase procedure. The entries of the matrix are first sampled uniformly at random to estimate, via a spectral method, the *leverage scores* of its rows and columns. These scores are then used to extract a few important rows and columns whose entries are further sampled. The algorithm exploits these new samples to complete the matrix estimation using a CUR-like method. For this leveraged matrix estimation procedure, we establish entry-wise guarantees that remarkably, do not depend on the coherence of the matrix but only on its spikiness. These guarantees imply that `LoRa-PI` learns an ε-optimal policy using $\widetilde{O}(\frac{S+A}{\mathrm{poly}(1-\gamma)\varepsilon^2})$ samples where S (resp. A) denotes the number of states (resp. actions) and γ the discount factor. Our algorithm achieves this order-optimal (in S, A and ε) sample complexity under milder conditions than those assumed in previously proposed approaches.

1 Introduction

Reinforcement Learning (RL) methods when applied to dynamical systems with large state and action spaces suffer from the curse of dimensionality. For example, learning an ε-optimal policy in tabular discounted Markov Decision Processes (MDPs) with S states and A actions requires a number of samples scaling at least as $\frac{SA}{(1-\gamma)^3\varepsilon^2}$ [17, 36]. Fortunately, many real-world systems exhibit a latent structure that if learnt and exploited could drastically improve the statistical efficiency of RL methods [25, 38]. In this paper, we are interested in developing methods to leverage low-rank latent structures. These structures have attracted a lot of attention recently, see e.g. [22, 11, 15, 28, 16, 45, 39, 2, 29, 40, 32, 35, 37, 34]. Here, we consider a structure where the (state, action) value functions of policies, viewed as $S \times A$ matrices, are low-rank. This structure has been empirically motivated and studied in [35, 34, 43, 33]. The hope is that when exploiting it optimally, learning an ε-optimal policy would only require $O(\frac{S+A}{(1-\gamma)^3\varepsilon^2})$ samples. Such an improvement would also imply significant statistical gains in MDPs with continuous state and action spaces. If these spaces are of dimensions d_1 and d_2, under natural smoothness conditions and using an appropriate discretization [35], the sample complexity would indeed be reduced from $\frac{1}{\varepsilon^{d_1+d_2+2}}$ (without structure) to $\frac{1}{\varepsilon^{\max(d_1,d_2)+2}}$.

18th European Workshop on Reinforcement Learning (EWRL 2025).

In this paper, we present LoRa-PI (Low Rank Policy Iteration), a model-free algorithm that learns and exploits an initially hidden low-rank structure in MDPs. Unlike existing algorithms, LoRa-PI does not require any prior information on the structure. Yet, the algorithm offers the promised statistical gains: its sample complexity essentially exhibits an order-optimal dependence in S, A and ε (i.e., $\frac{S+A}{\varepsilon^2}$).

Contributions. Our algorithm LoRa-PI relies on approximate policy iteration [4]. As such, it alternates between policy evaluation and policy improvement steps. The design and performance analysis of these two steps constitute our main contributions.

1. Leveraged matrix estimation with entry-wise guarantees. LoRa-PI sequentially updates a candidate policy whose (state, action) value function has to be estimated. This function can be seen as an $S \times A$ matrix that we consider to be low rank. The policy evaluation step then boils down to a novel low-rank matrix estimation procedure. We have two main constraints for this procedure. (i) To be sample efficient, the matrix should be estimated from noisy observations of only a few of its entries. (ii) For RL purposes (when integrated to LoRa-PI), the procedure should offer entry-wise performance guarantees. We present LME (Leveraged Matrix Estimation), a low-rank matrix estimation algorithm that meets these constraints. LME does not require knowledge of a priori unknown parameters of the matrix (such as its rank, condition number, spikiness, or coherence), and it is the first algorithm enjoying non-vacuous entry-wise guarantees even for coherent matrices.

Method	Err. Guarantees	Sampling	Assumption	Complexity
LME (ours)	entry-wise	adaptive	bounded spikiness	$\alpha^2(S+A)/\epsilon^2$
Algorithm 1 [35]	entry-wise	apriori fixed anchors	anchors apriori known	$\alpha^2(S+A)/\epsilon^2$
LR-EVI (Thm 9 [34])	entry-wise	unif. anchors	incoherence	$\mu^2\alpha^2(S+A)/\epsilon^2$
NNM [9] (Thm 21 [34])	entry-wise	unif. anchors	incoherence	$\mu^2\alpha^2(S+A)/\epsilon^2$
Two-phase MC [7]	exact recovery	adaptive	noiseless	not applicable

Table 1: Comparison of methods with entry-wise guarantees. For brevity, the factors $(1-\gamma)^{-1}$, κ and d are omitted. NNM: nuclear norm minimization, MC: matrix completion.

More precisely, LME guarantees an entry-wise estimation error within ε using only $\widetilde{O}\left(\kappa^4\alpha^2\frac{(S+A)+\alpha^2}{(1-\gamma)^3\varepsilon^2}\right)$ samples, where α and κ denote the spikiness and the condition number of the matrix, respectively. Note that in particular, this sample complexity does not depend on the coherence of the matrix. Its dependence in S, A and ε cannot be improved. To reach this level of performance, LME relies on an adaptive sampling strategy. It first estimates, via a spectral method, the so-called *leverage scores* of the matrix. These scores quantify the amount of information about the matrix available in the different rows and columns. The algorithm then exploits the leverage scores to adapt its strategy and in turn, drive the sampling process towards more informative entries.

2. Design and sample complexity of LoRa-PI. Our RL algorithm LoRa-PI is a policy iteration algorithm that relies on LME to perform policy evaluation steps. The algorithm inherits the advantages of LME. In contrast to existing algorithms, it is parameter-free and its performance can be analyzed and guaranteed under mild assumptions on the (state, actions) value functions. In particular, the corresponding low-rank matrices do not need to be incoherent. We establish that LoRa-PI learns an ε-optimal policy using $\widetilde{O}\left(\kappa^4\alpha^2\frac{(S+A)+\alpha^2}{(1-\gamma)^8\varepsilon^2}\right)$ samples, where α and κ are upper bounds on the spikiness and the condition number of the (state, action) value functions.

3. Numerical experiments. We illustrate numerically the performance of our algorithms, LME and LoRa-PI, using synthetically generated low-rank MDPs. The experiments are presented in Appendix A due to space constraints.

Notation. We denote the Euclidean norm of a vector x by $\|x\|_2$. Let M be an $m \times n$ matrix. We we denote its i-th row (resp. j-th column) by $M_{i,:}$ (resp. by $M_{:,j}$). We denote its operator norm by $\|M\|_{\mathrm{op}}$, it Frobenius norm by $\|M\|_{\mathrm{F}}$, its infinity norm by $\|M\|_\infty = \max_{i\in[m],j\in[n]} |M_{ij}|$, and its two-to-infinity norm by $\|M\|_{2\to\infty} = \max_{i\in m} \|M_{i,:}\|_2$. We denote by $M^\dagger$ the Moore-Penrose inverse of M. For given subsets $\mathcal{I} \in [m]$, $\mathcal{J} \in [n]$, we denote by $M_{\mathcal{I},\mathcal{J}}$ the sub-matrix whose entries are $\{M_{ij} : (i,j) \in \mathcal{I} \times \mathcal{J}\}$. Finally, we use $a \wedge b = \min(a,b)$ and $a \vee b = \max(a,b)$.

2 Related Work

Low-rank MDPs. MDPs with low-rank latent structure have been extensively studied recently. We may categorize these studies according to the type of the underlying low-rank structure and to the nature of the algorithms used to learn this structure.

The most studied low-rank structure concerns MDPs whose transition kernels and the expected reward functions are low-rank. For instance, it is assumed that the transition probabilities can be written as $p(s'|s,a) = \phi(s,a)^\top \mu(s')$, where $\phi(s,a)$ and $\mu(s')$ are d-dimensional feature maps [22, 11, 15, 28, 16, 45, 39, 2, 29, 40, 32]. These work additionally assume that the feature map ϕ (and similarly for μ) belongs to a rich function class $\mathcal{H}$. In this setting, the typical upper bounds derived for the sample complexity of identifying an ε-optimal policy scale as $\mathrm{poly}(A,(1-\gamma)^{-1})\frac{\log|\mathcal{H}|}{\varepsilon^2}$. When no restrictions are imposed on the class $\mathcal{H}$, one can find a low-rank structure such that $\log|\mathcal{H}|$ scales as the number S of states [20]. In this case, the aforementioned upper bounds are the same those for MDPs without structure. We also note that most algorithms using this framework rely on strong computational oracles (e.g., empirical risk minimizers, maximum likelihood estimators), see [23, 18, 44] for detailed discussions. In this paper, we do not limit our analysis to low-rank structures based on a given restricted class of functions, and our algorithms do not rely on any kind of oracle.

The low-rank structure we consider is similar to that in [35, 34] and just assumes that the (state, action) value functions are low-rank. Actually, [35] considers the case where only the optimal Q-function is low-rank, say of rank d. As shown in [35], such a structure naturally arises when discretizing smooth MDPs with continuous state and action spaces. In both papers [35, 34], the authors devise algorithms with a minimax-optimal sample complexity to identify an ε-optimal policy roughly scaling as $(S+A)/\varepsilon^2$. But the analysis presented in [35] suffers from the following important limitations. 1. First, it is assumed that the learner is aware of a set $\mathcal{I}$ (resp. $\mathcal{J}$) of so-called anchors states (resp. actions), such that the rank of the matrix $Q_{\mathcal{I},\mathcal{J}} := (Q(s,a))_{(s,a)\in\mathcal{I}\times\mathcal{J}}$ is the same as that of the entire matrix Q. Such anchors are however initially unknown (since Q is unknown). Importantly, the proposed RL algorithms rely on a low-rank matrix estimation procedure whose performance strongly depends on the smallest singular value $\sigma_d(Q_{\mathcal{I},\mathcal{J}})$ of $Q_{\mathcal{I},\mathcal{J}}$. The authors circumvent this difficulty by actually parametrizing their algorithms using $\sigma_d(Q_{\mathcal{I},\mathcal{J}})$. But again, the latter is unknown, and it remains unclear how one can avoid this issue. 2. The second limitation is that the analysis is valid for small values of the discount factor γ (the authors need to impose an upper bound on $\gamma/\sigma_d(Q_{\mathcal{I},\mathcal{J}})$). When $\sigma_d(Q_{\mathcal{I},\mathcal{J}})$ is small, the analysis is limited to very short horizons. Note that in addition, [35] assumes that the collected rewards are deterministic, which together with the short horizon issue, greatly simplifies the learning problem.

To address the first limitation, the authors of [34] propose to sample rows and columns uniformly at random to get anchors. This solution requires to sample at least $d\mu^2$ states and actions (Lemma 10 in [34]) where μ is the (unknown) coherence of the matrix to be estimated. Hence this essentially amounts to sampling almost the whole matrix for coherent matrices. The authors of [34] also propose a solution to the second limitation, but at the expense of imposing additional restrictive conditions. In this paper, we address both limitations and devise RL algorithms that rely on a new low-rank matrix estimation procedure that works without imposing the incoherence of the matrix and that does not require knowledge on a priori unknown parameters of this matrix.

Low-rank matrix estimation with entry-wise guarantees. Until recently, most results on low-rank matrix recovery concerned guarantees with respect to the spectral or Frobenius norms, see e.g. [12] and references therein. Over the past few years, methods to derive entry-wise guarantees have been developed. These include spectral approaches [1, 8, 37], nuclear-norm penalization and convex optimization techniques [9], CUR-based (or Nyström-like) methods [35, 3, 34].

The aforementioned literature provides guarantees not for all low-rank matrices, but for those typically enjoying additional structural properties such as incoherence. Relaxing the incoherence assumption is not easy, but can be achieved using adaptive sampling [24, 7, 41]. As far as we are aware, all results applicable to somewhat coherent matrices provide guarantees with respect to the spectral or Frobenius norms. In this paper, we develop a first adaptive matrix estimation method with provable entry-wise guarantees, valid for matrices with well-defined spikiness but not necessarily incoherent. Refer to [26, 30] and to §3.2 for a detailed discussion about the notion of spikiness.

3 Preliminaries

3.1 Low-rank Markov Decision Processes

We consider a discounted MDP with finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$. These spaces are of cardinality S and A, respectively. The dynamics are described by the transition kernel p where $p(s'|s,a)$ denotes the probability to move to state s' given current state s and that the action a is selected. The collected rewards are random but bounded by $r_{\max}$, and $r(s,a)$ is the expected reward collected when action a is selected in state s. A deterministic Markovian policy π is described by a mapping from $\mathcal{S}$ to $\mathcal{A}$. We denote by V^π the state value function of π: for all $s \in \mathcal{S}$, $V^\pi(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t^\pi, a_t^\pi) | s_0^\pi = s]$, where s_t^π and a_t^π are, at time t, the state and the action selected under π. Similarly, the (state, action) value function of π is defined by: for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $Q^\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^\pi(s')$. Q^π can be seen as a $S \times A$ matrix, referred to as the *value matrix of* π in the remainder of the paper. Let κ_π denote the condition number of Q^π. Finally, let $V^\star$ be the value function of the MDP (the value function of the optimal policy).

The objective is to learn an ε-optimal policy by interacting with the MDP. Such a policy satisfies: for all $s \in \mathcal{S}$, $V^\pi(s) \geq V^\star(s) - \varepsilon$. Without any assumption on the structure of the MDP, to identify such a policy, the learner needs to gather, even with a generative model, a number of samples[1] that scales as $\frac{SA}{\varepsilon^2(1-\gamma)^3}$ [17, 36]. The hope is that exploiting an a-priori known structure in the MDP may considerably accelerate the learning process. In this paper, we focus on a low-rank latent structure. Formally, we define:

Definition 1 (Rank of a policy, rank of the MDP)**.** *The rank d_π of a deterministic policy π is the rank of its value matrix Q^π. The rank of an MDP is then defined as $d = \max_\pi d_\pi$, where the maximum is over all deterministic policies.*

Throughout the paper, we assume that the MDP is low-rank: its rank d satisfies $d \ll (S + A)$. This assumption is merely made to simplify the exposition of our results and proof techniques. As we shall argue in Appendix E, our findings can naturally be extended to MDPs that are only low-rank in an approximate and well-precised sense.

3.2 Matrix estimation: coherence and spikiness

Our learning algorithm relies on the approximate policy iteration method, and in particular, in each iteration, it needs to estimate the low-rank value matrix of the current policy. To be sample efficient, the algorithm will estimate the matrix from the noisy observations of a few of its entries. Recovering a low-rank matrix from a few of its entries is not always possible (see e.g. [12] for a survey), and conditions on the degree to which information about a single entry is spread out across a matrix must be imposed. Examples of such conditions pertain to the *coherence* [6, 31] or the *spikiness* [30] of the matrix.

Matrix coherence. Let Q be a rank-d $S \times A$ matrix with SVD $U \Sigma W^\top$. The coherence of Q is defined as $\mu(Q) = \max\{\sqrt{S/d}\|U\|_{2\to\infty}, \sqrt{A/d}\|W\|_{2\to\infty}\}$. Q is μ-coherent if $\mu(Q) \leq \mu$. Low coherence means that the energy of U and W are not concentrated around a few rows and columns.

Matrix spikiness. The spikiness of Q is defined as $\alpha(Q) = \sqrt{SA}\|Q\|_\infty / \|Q\|_{\mathrm{F}} \in [1, \sqrt{SA}]$. Q is α-spiky if $\alpha(Q) \leq \alpha$. A matrix has low spikiness if the amplitude of its maximal entry is not much larger than the average amplitude of its entries, in which case, it is intuitively easier to estimate.

Most existing guarantees for low-rank matrix estimation are expressed through the spectral or Frobenius norm of the error matrix. For this type of guarantees, the estimation error scales polynomially either with the matrix coherence or with its spikiness [12, 30]. The matrix spikiness was introduced in the matrix completion literature [30] to obtain guarantees under less restrictive conditions than the incoherence conditions imposed in previous work. Indeed, there are matrices with bounded spikiness but high coherence (say close to $\sqrt{S/d}$, in which case the aforementioned coherence-based guarantees are vacuous). In contrast, bounded incoherence provides an upper bound on spikiness since $\alpha(Q) = \sqrt{SA}\|Q\|_\infty / \|Q\|_{\mathrm{F}} \leq \sqrt{SA}\|U\|_{2\to\infty}\|Q\|_{\mathrm{op}}\|W\|_{2\to\infty} / \|Q\|_{\mathrm{F}} \leq \mu(Q)^2 d$.

[1] Here a sample refers to an experience (s, a, r, s'), the observation of the collected reward r and the next state s', starting with a given (state, action) pair (s, a). Under a generative model, the learner can adapt the choice of (s, a) for the next observed experience without any constraint.

For RL purposes, we need to derive entry-wise guarantees for the estimate of the value matrix of some policy as demonstrated in [35, 37, 21]. Existing upper bounds for the entry-wise estimation error exhibit a strong dependence in the matrix coherence and its condition number, see e.g. [9, 8, 37]. For instance, in [9], this dependence comes as a multiplicative factor $\mu(Q)^2\alpha(Q)^2\kappa(Q)^2$ in the number of samples required for a given level of estimation accuracy. As far as we are aware, our matrix estimation method is the first able to yield entry-wise guarantees that do not exhibit a dependence on the matrix coherence but only on its spikiness (see Table 1). Our algorithm is better by a factor of $\mu(Q)^2$ than algorithms based on uniform sampling (studied in [34]), and requires the same sample complexity as the algorithm of [35], which has prior knowledge of anchor states. It remains however unclear whether the dependence of the entry-wise estimation error in the condition number can be avoided. This last observation guides the design of RL algorithms for low-rank MDPs as we discuss next.

3.3 Policy vs. Value Iteration: the condition number issue

We aim at devising an algorithm learning an efficient policy with provable guarantees while imposing conditions on the MDP that are as mild as possible. To this aim, one may think of applying either a policy iteration approach, as we do, or a value iteration approach.

Policy Iteration. Using this approach, in each iteration, we need to estimate the low-rank value matrix of the current candidate policy. As mentioned above, the entry-wise error of this estimation procedure depends on the condition number of the matrix. Note that this matrix belongs to the finite set of (state, action) value functions of deterministic policies. As shown in [10, 42], this set can be seen as the vertices of a simple polytope $\mathcal{P}$. Hence to get performance guarantees when applying a PI approach, it is sufficient to impose an upper bound on the condition numbers κ_π for all deterministic policies π, or equivalently, on the condition numbers of matrices corresponding to the vertices of $\mathcal{P}$.

Value Iteration. Here, we would maintain, in iteration t, an estimate $V^{(t)}$ of the value function $V^\star$, and samples would be used to compute $V^{(t+1)}$, an estimate of $\mathcal{T}^\star(V^{(t)})$, where $\mathcal{T}^\star$ denotes Bellman's operator. More precisely, starting from $V^{(t)}$, we would estimate the low-rank matrix $Q^{(t+1)} = \mathcal{F}(V^{(t)})$ defined by for all (s,a), $\mathcal{F}(V^{(t)})(s,a) = r(s,a) + \gamma\sum_{s'} p(s'|s,a)V^{(t)}(s')$. Then we would define $V^{(t+1)}$ as the value function of the greedy policy with respect to $Q^{(t+1)}$. Hence to get provable performance guarantees using a value iteration approach, we would need to impose an upper bound on the condition number of $Q^{(t)}$ in all iterations t. The main issue is that the set of matrices $\{Q^{(t)}, t \geq 1\}$ is stochastic and hard to predict. Indeed, we have no way of confining the iterates $Q^{(t+1)}$ to the polytope $\mathcal{P}$: as shown in [10], the polytope is not stable by Bellman's operator. As a consequence, if we wish to get performance guarantees for a value iteration approach, we would need to impose an upper bound on the condition number of all possible matrices of the form $\mathcal{F}(V)$ for some vector V.

In summary, policy iteration approaches offer a theoretical advantage compared to value iteration. It requires the control of the condition numbers of matrices in a set much smaller than that for value iteration. This advantage is illustrated in Figure 1 on a toy example of an MDP. Refer to Appendix A for additional numerical experiments (with larger MDPs).

4 Leveraged Matrix Estimation

In this section, we present Leveraged Matrix Estimation (`LME`), an algorithm that estimates the value matrix Q^π of a policy π. The algorithm relies on an active strategy for sampling the entries of the matrix based on its estimated leverage scores as defined below. This active strategy accelerates the learning process and allows us to obtain entry-wise guarantees that do not depend on the coherence of the matrix but on its spikiness only.

Definition 2 (Leverage scores[2])**.** *Let Q be a rank-d $S \times A$ matrix with SVD $U\Sigma W^\top$. Its left and right leverage scores ℓ and ρ are defined as $\ell_s = \|U_{s,:}\|_2^2/d$ for all $s \in \mathcal{S}$, and $\rho_a = \|W_{a,:}\|_2^2/d$ for all $a \in \mathcal{A}$.*

`LME` only takes as inputs a policy π and a sampling budget T. It proceeds in two phases: first, it uses half of the sampling budget to estimate the leverage scores of Q^π via singular subspace recovery.

[2]Our definition of leverage scores is consistent up to a scale factor with that used in the literature [5, 13, 7].

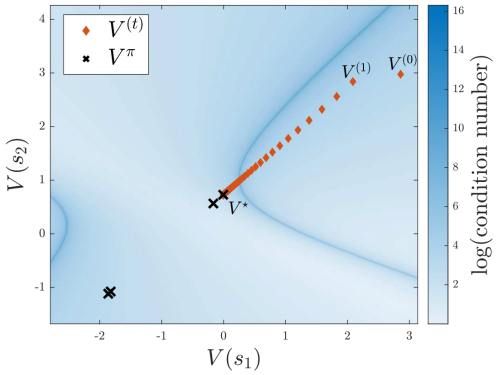


Figure 1: Consider an MDP with two states and two actions (see Appendix A.1 for details). The 4 black crosses correspond to the value function of the 4 possible policies. When combining policy iteration with a low rank estimation procedure, we just need to control the condition number of the 4 corresponding value matrices. The red dots correspond to the successive estimates $V^{(t)}$ of $V^{\star}$ when running value iteration. When applying a value iteration approach, we would need to upper bound the condition number of all the corresponding matrices $Q^{(t)} = \mathcal{F}(V^{(t-1)})$ for $t \geq 1$. For a given V, the background color in the figure indicates the value of the condition number of $\mathcal{F}(V)$. We see that the dynamics of $V^{(t)}$ under the value iteration algorithm are such that the trajectory $(Q^{(t)}, t \geq 1)$ has to go through regions where the condition number is very high. Hence on this example, a value iteration approach would not work well.

Second, it selects a few anchor rows and columns sampled using the estimated leverage scores, and uses the remaining budget to sample the entries of these rows and columns. It finally completes the matrix estimation using a CUR-based method. The full pseudo-code of `LME` is presented in Appendix C. Observe that `LME` is parameter-free: it does not require knowledge of the policy rank d_π, nor upper bounds on unknown parameters such as κ_π or $\alpha(Q^\pi)$ or $\mu(Q^\pi)$. Throughout this section, when presenting our guarantees, we will abuse notation and use d, κ and α, instead of d_π, κ_π and $\alpha(Q^\pi)$.

4.1 Preliminaries

`LME` exploits a natural empirical estimator of Q^π entries at numerous stages. This empirical estimator is essentially based on Monte-Carlo rollouts with truncation as described next. Define the truncated value matrix at a horizon τ as follows: for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $Q^\pi_\tau(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t(s_t^\pi, a_t^\pi) \mathbb{1}_{\{t \leq \tau\}} \middle| s_0^\pi = s, a_0^\pi = a\right]$. By choosing τ appropriately, we may control the level of the approximation error $Q^\pi_\tau - Q^\pi$. We make this observation precise in the following lemma, proved in Appendix F.1.

Lemma 1. *For any $\epsilon > 0$ and any $\tau \geq \frac{1}{1-\gamma} \log\left(\frac{r_{\max}}{(1-\gamma)\epsilon}\right)$, we have $\|Q^\pi - Q^\pi_\tau\|_\infty \leq \epsilon$.*

In view of the above, to estimate an entry, say (s,a), of Q^π, we will use an empirical estimator based on trajectories of length $\tau + 1$ of the system under π and starting with (state, action) pair (s,a). In our algorithms, this length is chosen to get an appropriate accuracy level. Specifically, we choose ϵ and τ as follows:

$$\epsilon = \frac{r_{\max}}{T} \quad \text{and} \quad \tau = \left\lceil \frac{1}{1-\gamma} \log\left(\frac{T}{1-\gamma}\right) \right\rceil. \tag{1}$$

These choices will become apparent from our analysis.

4.2 Phase 1: Leverage scores estimation via spectral subspace recovery

The first phase of `LME` is devoted to the estimation of the leverage scores of Q^π. To this aim, using half of the sampling budget $T/2$, we estimate the singular subspaces of the matrix via a spectral method.

Phase 1a. Data collection and the empirical truncated value matrix. As suggested in §4.1, to estimate individual entries of Q^π, we sample system trajectories of length $\tau+1$. More precisely, for each of the $N := T/(2(\tau+1))$ trajectories, we first sample the starting (state, action) pair uniformly at random, and then observe the trajectory obtained under the policy π and initiated at this pair. The data collected this way is $\mathcal{D} = \{(s^\pi_{k,0}, a^\pi_{k,0}, r^\pi_{k,0}, \dots, s^\pi_{k,\tau}, a^\pi_{k,\tau}, r^\pi_{k,\tau}) : k \in [N]\}$. Using this data, we construct an empirical estimate of the truncated value matrix as follows $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$:

$$\widetilde{Q}^\pi_\tau(s,a) = \frac{SA}{N}\sum_{k=1}^{N}\left(\sum_{t=0}^{\tau}\gamma^t r^\pi_{k,t}\right)\mathbb{1}\{(s^\pi_{k,0}, a^\pi_{k,0}) = (s,a)\}, \tag{2}$$

Phase 1b. Singular subspace recovery. We compute the SVD of the empirical truncated value matrix $\widetilde{Q}^\pi_\tau$. We obtain $\widetilde{Q}^\pi_\tau = \sum_{i=1}^{S\wedge A}\hat{\sigma}_i\hat{u}_i\hat{w}_i^\top$, where $\hat{\sigma}_1,\dots,\hat{\sigma}_{S\wedge A}$ correspond, in decreasing order, to its singular values and $\hat{u}_1,\dots,\hat{u}_S$ (resp. $\hat{w}_1,\dots,\hat{w}_A$) to its left (resp. right) singular vectors. Using this decomposition, we construct our estimate of Q^π as follows:

$$\widehat{Q}^\pi = \sum_{i=1}^{S\wedge A}\hat{\sigma}_i\mathbb{1}\{\hat{\sigma}_i \ge \beta\}\hat{u}_i\hat{w}_i^\top, \tag{3}$$

where $\beta > 0$ is a threshold that we will precise shortly. We view $\widehat{Q}^\pi$ as a biased estimate of Q^π with controlled bias through τ. We also use β to estimate the rank of Q^π: $\widehat{d} = \sum_{i=1}^{S\wedge A}\mathbb{1}\{\hat{\sigma}_i \ge \beta\}$. Finally, the estimated left (resp. right) singular subspace is denoted $\widehat{U} = [\hat{u}_1 \quad \cdots \quad \hat{u}_{\hat{d}}] \in \mathbb{R}^{S\times\hat{d}}$ (resp. $\widehat{W} = [\hat{w}_1 \quad \cdots \quad \hat{w}_{\hat{d}}] \in \mathbb{R}^{A\times\hat{d}}$). In the following proposition, we provide a choice for the threshold β that yields appropriate guarantees regarding our subspace recovery.

Proposition 1. *Let $\delta \in (0,1)$ and choose the threshold β as*

$$\beta = \sqrt{\frac{r^2_{\max}SA(S+A)}{(1-\gamma)^3 T}\log^4\left(\frac{(S+A)T}{(1-\gamma)\delta}\right)} + \frac{r_{\max}\sqrt{SA}}{T}. \tag{4}$$

Then, provided that[3]*:*

$$T = \widetilde{\Omega}_\delta\left(\frac{r^2_{\max}SA}{\sigma^2_d(Q^\pi)}\frac{(S+A)}{(1-\gamma)^3}\right) \tag{5}$$

we have that events: $\widehat{d} = d_\pi$, and for all $s \in \mathcal{S}$,

$$\|U_{s,:} - \widehat{U}_{s,:}(\widehat{U}^\top U)\|_2 \lesssim \frac{r_{\max}\sqrt{SA}}{(1-\gamma)^{3/2}\sigma_d(Q^\pi)}\left(\sqrt{\frac{d}{T}} + \kappa\|U_{s,:}\|_2\sqrt{\frac{S+A}{T}}\right)\log^2\left(\frac{(S+A)T}{(1-\gamma)\delta}\right)$$

hold with probability at least $1-\delta$. An analogous result holds for $\widehat{W}$.

The precise statement (Theorem 4) and the proof are presented in Appendix B.3 and B.4.

Phase 1c. Leverage Scores Estimation. To conclude, using the recovered subspaces $\widehat{U}$ and $\widehat{W}$, we estimate the leverage scores as follows $\hat{\ell} = \|\tilde{\ell}\|_1^{-1}\tilde{\ell}$ and $\hat{\rho} = \|\tilde{\rho}\|_1^{-1}\tilde{\rho}$, where:

$$\forall s \in \mathcal{S}:\ \tilde{\ell}_s = \|\widehat{U}_{s,:}\|_2^2 \vee \frac{d}{S}, \qquad \text{and} \qquad \forall a \in \mathcal{A}:\ \tilde{\rho}_a = \|\widehat{W}_{a,:}\|_2^2 \vee \frac{d}{A}. \tag{6}$$

The performance of the estimation of the leverage scores is summarized in the following theorem, proved in Appendix B.2.

Theorem 1 (Leverage Scores Estimation)**.** *Let $\delta \in (0,1)$. Suppose the threshold β is chosen as in (4). Then, we have that: $\mathbb{P}(\forall s \in \mathcal{S},\ \ell_s \le 4\,\hat{\ell}_s) \ge 1-\delta$, provided that*

$$T = \widetilde{\Omega}_\delta\left(\kappa^2\frac{r^2_{\max}SA}{\sigma^2_d(Q^\pi)}\frac{(S+A)}{(1-\gamma)^3}\right),$$

An analogous result holds for $\hat{\rho}$.

[3]To simplify the notation, all our sample complexity guarantees are expressed using $\widetilde{\Omega}_\delta(\cdot)$, the tilde-notation may hide poly-log dependencies in δ, S, A, $(1-\gamma)^{-1}$, d, κ, α, $\log(e/\varepsilon)$, and $r_{\max}$.

4.3 Phase 2: Leveraged CUR-based Matrix Completion

Before we proceed with the description of the second phase, we briefly recall the so-called CUR decomposition [19, 27] for low-rank matrices. The decomposition says that for a given rank-d $S \times A$ matrix Q, there always exists $\mathcal{I} \subseteq [S]$, $\mathcal{J} \subseteq [A]$, with $|\mathcal{I}| = |\mathcal{J}| = d$, such that the sub-matrix $Q_{\mathcal{I},\mathcal{J}}$ is full rank and for all entries (i,j), $Q_{ij} = Q_{i,\mathcal{J}}(Q_{\mathcal{I},\mathcal{J}})^{\dagger}Q_{\mathcal{I},j}$. As in [35, 34, 3], we leverage this decomposition in our matrix estimation procedure, but without any requirement such as knowledge of $\mathcal{I}, \mathcal{J}$ for which $\sigma_d(Q_{\mathcal{I},\mathcal{J}})$ bounded away from zero or upper bounds on parameters like the matrix coherence.

Phase 2a. Data collection to estimate the skeleton of the value matrix. We start by sampling $K := 64d\log(64d/\delta)$ rows (resp. columns) without replacement according to $\hat{\ell}$ (resp. $\hat{\rho}$) to form a *skeleton* of the matrix. These rows and columns are referred to as anchors. We denote the set of selected rows (resp. columns) by $\mathcal{I} \subseteq \mathcal{S}$ (resp. $\mathcal{J} \subseteq \mathcal{A}$). We use the remaining sample budget $T/2$ to get samples of the entries of Q^{π} in the skeleton. To this aim, we use the procedure described in §4.1, and sample trajectories of length $\tau + 1$. For each entry $(s,a) \in \Omega_{\square} := \mathcal{I} \times \mathcal{J}$, we use $N_1 := T/(4(\tau+1)K^2)$ trajectories to compute $\widetilde{Q}^{\pi}_{\tau}(s,a)$, an empirical estimate of $Q^{\pi}(s,a)$ (see (2)). For each entry $(s,a) \in \Omega_{+} := ((\mathcal{S}\backslash\mathcal{I})\times\mathcal{J}) \cup (\mathcal{I}\times(\mathcal{A}\backslash\mathcal{J}))$, we use $N_2 := T/(4(\tau+1)(K(S+A)-2K^2)$ trajectories. Note that $N_2 \leq N_1$ (this plays a role in the analysis).

Phase 2b. CUR-based completion with Inverse Leverage Scores Weighting. First, using the leverage scores, and the set of rows $\mathcal{I}$ and columns $\mathcal{J}$, we define $K \times K$ diagonal matrices L and R as follows:

$$\forall i \in \mathcal{I}, \quad L_{ii} = \frac{1}{\min\left\{1, \sqrt{K\hat{\ell}_i}\right\}}, \quad \text{and} \quad \forall j \in \mathcal{J}, \quad R_{jj} = \frac{1}{\min\{1, \sqrt{K\hat{\rho}_j}\}}. \tag{7}$$

Next, starting from the values of $\widetilde{Q}^{\pi}_{\tau}(s,a)$ for (s,a) in the skeleton, we perform a CUR matrix completion to obtain $\widehat{Q}^{\pi}$: *(i)* for all $(s,a) \in (\mathcal{S}\times\mathcal{J}) \cup (\mathcal{I}\times\mathcal{A})$, we set $\widehat{Q}^{\pi}(s,a) = \widetilde{Q}^{\pi}_{\tau}(s,a)$; *(ii)* for all $(s,a) \in (\mathcal{S}\backslash\mathcal{I}) \times (\mathcal{A}\backslash\mathcal{J})$, we set

$$\widehat{Q}^{\pi}(s,a) = \widetilde{Q}^{\pi}_{\tau}(s,\mathcal{J})R\left(L\,\widetilde{Q}^{\pi}_{\tau}(\mathcal{I},\mathcal{J})R\right)^{\dagger}L\,\widetilde{Q}^{\pi}_{\tau}(\mathcal{I},a). \tag{8}$$

Note that the use of L and R in (8), referred to as Inverse Leverage Scores Weighting, corresponds to an importance sampling procedure. It allows us to account for the fact that the skeleton has been sampled using the (estimated) leverage scores.

The next theorem summarizes the performance guarantees under `LME`. Its proof is presented in Appendix C.1.

Theorem 2. *Let $\varepsilon > 0$, $\delta \in (0,1)$. Given a deterministic policy π, and a sampling budget T, the algorithm* `LME` *ensures that $\mathbb{P}(\|\widehat{Q}^{\pi} - Q^{\pi}\|_{\infty} \leq \varepsilon) \geq 1-\delta$, provided that $\varepsilon \lesssim \|Q^{\pi}\|_{\infty}$ and*

$$T = \widetilde{\Omega}_{\delta}\left(\frac{(S+A)+\alpha^2 d}{(1-\gamma)^3\varepsilon^2}(r^2_{\max}\kappa^4\alpha^2 d^2)\right).$$

Theorem 2 states that the sample complexity of `LME` to obtain entry-wise guarantees does not depend on the coherence μ of Q^{π} but rather on its spikiness α and condition number κ only. Hence `LME` provides entry-wise guarantees even for coherent matrices. In addition, its sample complexity scales with S, A, γ and ε optimally. Indeed if $\alpha, \kappa = \Theta(1)$ and $d \ll S + A$, it scales as $\frac{(S+A)}{\varepsilon^2(1-\gamma)^3}$. We also wish to emphasize that `LME` is parameter-free, in the sense that it does not require knowledge of the so-called anchor rows and columns, nor does it require upper bounds on unknown parameters such as coherence, spikiness, rank or condition number. These properties are desirable for RL purposes.

5 Low-Rank Policy Iteration

In this section, we present and evaluate `LoRa-PI` (Low Rank Policy Iteration), a model-free variant of the approximate policy iteration algorithm [4]. It alternates between policy improvement and policy evaluation steps and uses `LME`, our low rank matrix estimation procedure for policy evaluation. Refer to Algorithm 1 for the pseudo-code.

Algorithm 1: Low-Rank Policy Iteration (`LoRa-PI`)

Input: sampling budget T, accuracy ε, confidence level δ, discount factor γ, rewards upper bound $r_{\max}$, $\pi^{(1)}$ an initial deterministic policy

Set $N_{\text{epochs}} \leftarrow (1-\gamma)^{-1}\log\left((4r_{\max})/((1-\gamma)\varepsilon)\right)$, $T_{\text{eval}} \leftarrow T/N_{\text{epochs}}$

for $t = 1, 2, \ldots, N_{\text{epochs}}$ **do**

(Approximate policy evaluation). Obtain the value matrix of $\pi^{(t)}$: $\widehat{Q}^{(t)} \leftarrow \texttt{LME}(\pi^{(t)}, T_{\text{eval}})$

(Policy improvement) Improve $\pi^{(t)}$ using $\widehat{Q}^{(t)}$: $\forall s \in \mathcal{S}, \pi^{(t+1)}(s) \leftarrow \arg\max_{a\in\mathcal{A}} \widehat{Q}^{(t)}(s,a)$

end

Output: Output $\hat{\pi} = \pi^{(N_{\text{epochs}}+1)}$.

The following theorem provides performance guarantees for `LoRa-PI`. We state the results under the assumption that for any deterministic policy π, Q^π is α-spiky and has a condition number upper bounded by κ. The proof of Theorem 3 is presented in Appendix D.

Theorem 3. *Let* $\delta \in (0,1)$ *and* $\varepsilon = \widetilde{O}(\|Q^{\pi^{(1)}}\|_\infty)$. *Under* `LoRa-PI`, *we have* $\mathbb{P}\left(\|V^\star - V^{\hat{\pi}}\|_\infty \le \varepsilon\right) \ge 1-\delta$, *provided*

$$T = \widetilde{\Omega}_\delta\left(\frac{(S+A)+\alpha^2 d}{(1-\gamma)^8\varepsilon^2}(r_{\max}^2\kappa^4\alpha^2 d^2)\right).$$

`LoRa-PI` combines numerous advantages. (i) It is parameter-free: it does not require the knowledge of upper bounds on parameters such as the ranks, condition numbers, and spikiness of the value matrices of policies. This is thanks to `LME`, which is itself parameter-free. (ii) Its sample complexity does not depend on the coherence of the value matrices but only on their spikiness; which is an important improvement over existing algorithms [34]. (iii) `LoRa-PI` offers performance guarantees without having access to good anchor states and actions, without assuming that the rewards are deterministic and that the discount factor is (far too) small, as in [35] (refer to Section 2 for a detailed discussion). (iv) Its sample complexity has an order-optimal scaling in S, A and ε. (v) Finally, since `LoRa-PI` uses policy iteration, its theoretical guarantees can be established under milder assumptions than if value iteration was used instead (see §3.3).

The dependence of order $(1-\gamma)^{-8}$ is far from the ideal minimal dependence of order $(1-\gamma)^{-3}$ that one would typically obtain in RL without low-rank structure. This is an artifact of using a model-free approach, and more specifically the Monte-Carlo estimator of entries of the value matrices. Avoiding such high dependence requires further assumptions and a model-based approach. Furthermore, it is worth mentioning that the guarantees enjoyed by `LoRa-PI` can be naturally extended to MDPs that are low-rank only in an approximate sense. We refer the reader to Appendix E for further details.

6 Conclusion

In this work, we considered a class of MDPs where the Q-function, viewed as a state-action matrix, admits a low-rank representation under any deterministic policy. We devised `LoRa-PI`, a model-free learning algorithm based on approximate policy iteration, that provably exploits such low-rank representation to output a near-optimal policy. Critical to the design and performance guarantee of `LoRa-PI` is a novel low-rank matrix estimation procedure referred to as `LME`. `LME` is shown to enjoy a tight entry-wise guarantee while being parameter-free, i.e., it does not require knowledge of the so-called anchor rows and columns, nor upper bounds on unknown parameters such as spikiness, coherence, rank, or condition number. More importantly, its sample complexity does not scale with the coherence but instead with the spikiness of the matrix. This allows us to estimate a wider class of low-rank matrices with entry-wise guarantees than previous work. Such desirable properties are what make `LME` appealing for RL purposes, and in particular what allows us to show that `LoRa-PI` is sample-efficient under mild conditions. From a design perspective, `LME` and its analysis features many interesting tools and ideas. Notably, (i) we derived instance-dependent row-wise singular subspace recovery guarantees, and (ii) we combined the use of the so-called leverage scores with a CUR-based approximation for matrix estimation. We believe such tools and ideas to be of independent interest. Finally, we provided experimental results that suggest the superior performance of our proposed algorithms.

Acknowledgment

This research was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, the Swedish Research Council (VR), and Digital Futures. YJ is supported by the Knut and Alice Wallenberg Foundation Postdoctoral Scholarship Program under grant KAW 2022.0366.

References

[1] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452, 2020.

[2] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural Complexity and Representation Learning of Low Rank MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pages 20095–20107. Curran Associates, Inc., 2020.

[3] Anish Agarwal, Munther Dahleh, Devavrat Shah, and Dennis Shen. Causal matrix completion. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3821–3826. PMLR, 2023.

[4] Dimitri Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[5] Christos Boutsidis, Michael W Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 968–977. SIAM, 2009.

[6] Emmanuel J. Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[7] Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Completing any low-rank matrix, provably. *The Journal of Machine Learning Research*, 16(1):2999–3034, 2015.

[8] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

[9] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121, 2020.

[10] Robert Dadashi, Adrien Ali Taiga, Nicolas Le Roux, Dale Schuurmans, and Marc G Bellemare. The value function polytope in reinforcement learning. In *International Conference on Machine Learning*, pages 1486–1495. PMLR, 2019.

[11] Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On Oracle-Efficient PAC RL with Rich Observations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[12] Mark A. Davenport and Justin K. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE J. Sel. Top. Signal Process.*, 10(4):608–622, 2016.

[13] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13(1):3475–3506, dec 2012.

[14] Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

[15] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with Rich Observations via Latent State Decoding. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1665–1674. PMLR, 09–15 Jun 2019.

[16] Dylan Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning: A Disagreement-Based Perspective. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2059–2059. PMLR, 15–19 Aug 2021.

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

[18] Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Learning in Observable POMDPs, without Computationally Intractable Oracles. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

[19] Sergei A Goreinov, Eugene E Tyrtyshnikov, and Nickolai L Zamarashkin. A theory of pseudoskeleton approximations. *Linear algebra and its applications*, 261(1-3):1–21, 1997.

[20] Yassir Jedra, Junghyun Lee, Alexandre Proutiere, and Se-Young Yun. Nearly optimal latent state decoding in block mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 2805–2904. PMLR, 2023.

[21] Yassir Jedra, William Réveillard, Stefan Stojanovic, and Alexandre Proutiere. Low-rank bandits via tight two-to-infinity singular subspace recovery. In *Forty-first International Conference on Machine Learning*, 2024.

[22] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR, 06–11 Aug 2017.

[23] Daniel Kane, Sihan Liu, Shachar Lovett, and Gaurav Mahajan. Computational-statistical gap in reinforcement learning. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1282–1302. PMLR, 02–05 Jul 2022.

[24] Akshay Krishnamurthy and Aarti Singh. Low-rank matrix and tensor completion via adaptive sampling. *Advances in neural information processing systems*, 26, 2013.

[25] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. CURL: Contrastive Unsupervised Representations for Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5639–5650. PMLR, 13–18 Jul 2020.

[26] Lester W Mackey, Ameet Talwalkar, and Michael I Jordan. Distributed matrix completion and robust factorization. *J. Mach. Learn. Res.*, 16(1):913–960, 2015.

[27] Michael W Mahoney and Petros Drineas. Cur matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[28] Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic State Abstraction and Provably Efficient Rich-Observation Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6961–6971. PMLR, 13–18 Jul 2020.

[29] Aditya Modi, Jinglin Chen, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. Model-free representation learning and exploration in low-rank mdps. *Journal of Machine Learning Research*, 25(6):1–76, 2024.

[30] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13:1665–1697, 2012.

[31] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

[32] Tongzheng Ren, Tianjun Zhang, Lisa Lee, Joseph E Gonzalez, Dale Schuurmans, and Bo Dai. Spectral decomposition representation for reinforcement learning. In *Proc. of ICLR*, 2023.

[33] Sergio Rozada, Santiago Paternain, and Antonio G. Marques. Tensor and matrix low-rank value-function approximation in reinforcement learning. *IEEE Transactions on Signal Processing*, 72:1634–1649, 2024.

[34] Tyler Sam, Yudong Chen, and Christina Lee Yu. Overcoming the long horizon barrier for sample-efficient reinforcement learning with latent low-rank structure. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(2):1–60, 2023.

[35] Devavrat Shah, Dogyoon Song, Zhi Xu, and Yuzhe Yang. Sample efficient reinforcement learning via low-rank matrix estimation. *Advances in Neural Information Processing Systems*, 33:12092–12103, 2020.

[36] Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[37] Stefan Stojanovic, Yassir Jedra, and Alexandre Proutiere. Spectral entry-wise matrix estimation for low-rank reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling Representation Learning from Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9870–9879. PMLR, 18–24 Jul 2021.

[39] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in Contextual Decision Processes: PAC bounds and Exponential Improvements over Model-free Approaches. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2898–2933. PMLR, 25–28 Jun 2019.

[40] Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation Learning for Online and Offline RL in Low-rank MDPs. In *International Conference on Learning Representations*, 2022.

[41] Yining Wang and Aarti Singh. Provably correct algorithms for matrix column subset selection with selectively sampled data. *Journal of Machine Learning Research*, 18(156):1–42, 2018.

[42] Yue Wu and Jesús A. De Loera. Geometric policy iteration for markov decision processes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2070–2078, New York, NY, USA, 2022. Association for Computing Machinery.

[43] Yuzhe Yang, Guo Zhang, Zhi Xu, and Dina Katabi. Harnessing structures for value-based planning and reinforcement learning. In *International Conference on Learning Representations*, 2020.

[44] Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. Making Linear MDPs Practical via Contrastive Representation Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26447–26466. PMLR, 17–23 Jul 2022.

[45] Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient Reinforcement Learning in Block MDPs: A Model-free Representation Learning Approach. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26517–26547. PMLR, 17–23 Jul 2022.

A Numerical Experiments

All experiments in this section were performed on HP EliteBook 830 G8 with an Intel i7 core and 16 GB of RAM. Each experiment's runtime for individual realizations took at most 2-3 hours, and reproducing all results is feasible within a day.

A.1 Parameters of the toy example in Figure 1

We considered an MDP with $S = A = 2$, $\gamma = 0.87$, a reward matrix given by

$$r = \begin{bmatrix} -0.46 & -0.48 \\ -0.14 & 0.28 \end{bmatrix},$$

and the following transition probabilities:

$$P(s'|s, a = a_1) = \begin{bmatrix} 0.4 & 0.6 \\ 0.15 & 0.85 \end{bmatrix} \qquad P(s'|s, a = a_2) = \begin{bmatrix} 0.25 & 0.75 \\ 0.29 & 0.71 \end{bmatrix}$$

We initialized VI with $V^{(0)} = [2.86 \quad 2.98]^\top$. Note that $V_{\max} = \frac{r_{\max}}{1-\gamma} = 3.69$ and thus $V^{(0)} \in [-V_{\max}, V_{\max}]^2$.

For this example, the condition numbers of the Q-functions induced by policies are $16.08, 4.38, 15.29, 12.07$, while the maximum condition number during value iteration is ≈ 2497.82.

We stress here that this MDP is full-rank, and the purpose of this example is to demonstrate the potential instability of VI in the presence of large condition numbers. For low-rank MDPs, this corresponds to the matrix Q^π having an effectively smaller rank than expected, and estimating all d singular vectors despite $\sigma_1(Q^\pi)/\sigma_d(Q^\pi) \to \infty$.

A.2 Matrix completion with leveraged anchors

We consider matrix completion with a fixed matrix $M^\star$ to be estimated, testing four different methods. First, we test a method based on CUR-approximation with anchors chosen uniformly at random. Next, we have a method based on the estimation of leverage scores, where, for a given budget of samples, we use half of them for estimating leverage scores as described in the main text. Then, we consider a method with oracle anchors, where the anchors are chosen with respect to the true leverage scores. Lastly, we consider standard SVD decomposition, where we keep only the first d largest singular values of the matrix.

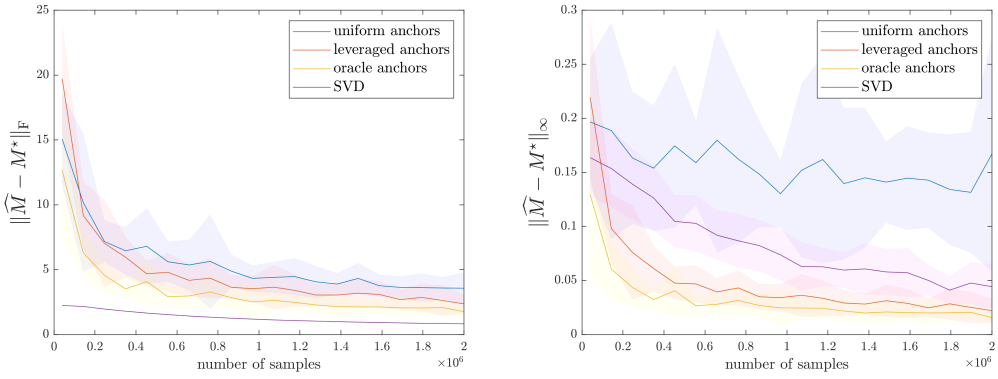
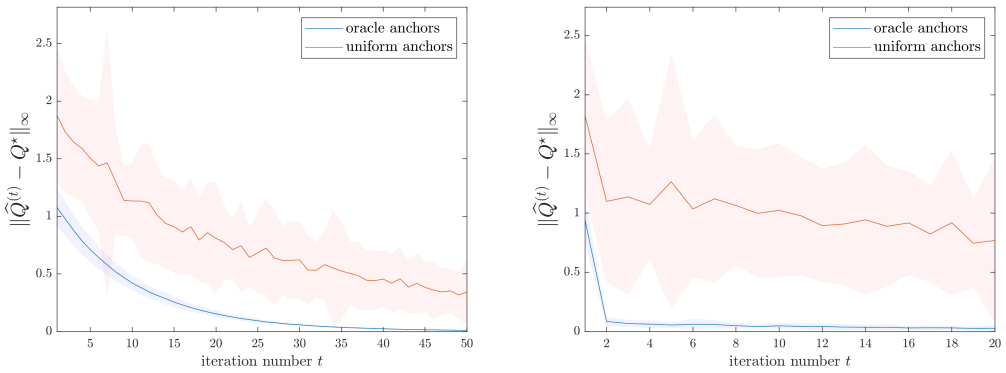


Figure 2: Matrix completion: matrix $M^\star$ is of size 1000×1000, rank $d = 5$ and sampled entries have additive Gaussian noise with $\sigma = 0.01$. Number of anchors used was $K = 10$. All plots are averaged over 30 simulations and a new random matrix $M^\star$ was generated in every 5 simulations.

As expected, CUR-based methods depend heavily on the quality of anchor selection. The gap between leverage-score-based anchors and oracle anchors is slight, even when half the samples are used to estimate the leverage scores. While SVD shows a smaller Frobenius error, it has higher entrywise error compared to CUR-based methods with good anchors.

A.3 Leverage scores for VI and PI

We demonstrate the importance of choosing anchors based on leverage scores for value iteration (VI) and policy iteration (PI). We postpone learning of the anchor states to the next subsections and assume that the true leverage scores of matrices $(Q^{(t)})_{t\geq 1}$ are given. For methods with leveraged anchors, anchors are chosen as those with the highest leverage scores (true leverage scores of $Q^{(t)}$). For uniform anchors, anchors are chosen uniformly without repetitions.



(a) Value iteration with anchors.

(b) Policy iteration with anchors.

Figure 3: Matrix $Q^\star$ is obtained from rank $d = 5$ rewards and transition matrices. Moreover, $S = 70, A = 50, \gamma = 0.9$, and we choose number of anchors $K = 15$. Observations are noisy with additive Gaussian noise with $\sigma = 0.01$. Plots are averaged over 100 simulations, and new MDPs are generated every 5 simulations, while the number of samples in an iteration t is $10(1.1)^t$.

These results highlight that leveraged anchors reduce entrywise error significantly for general matrices. In contrast, uniform anchors show significant randomness, although the error decreases in expectation over iterations.

A.4 Low-rank Value Iteration

We evaluate a VI-based variant of Algorithm 1, that we refer to as `LoRa-VI`. We do not assume prior knowledge of the matrices, and use samples to estimate leverage scores and matrices $(Q^{(t)})_{t\geq 1}$.

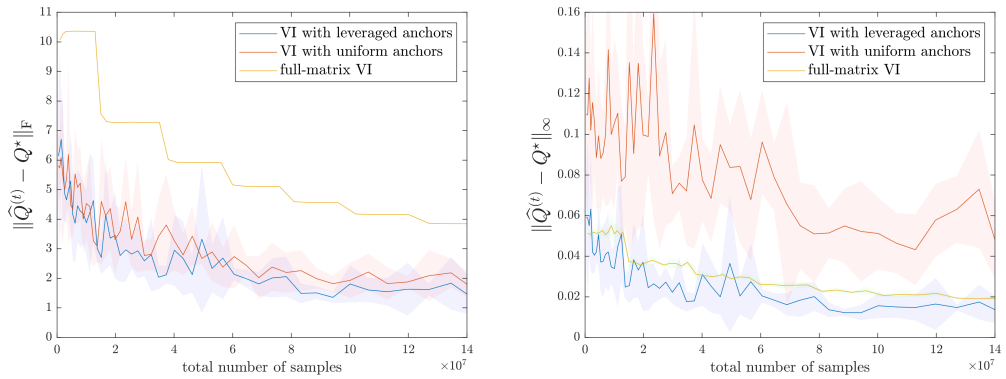


Figure 4: `LoRa-VI`: $Q^\star$ generated from low-rank r and P of rank $d = 4$, $S = A = 1000$, $\gamma = 0.1$. We used $K = 10$ anchors, $V^{(0)} = 0$, rewards are noisy with Gaussian noise $\sigma = 0.01$. All plots are averaged over 5 simulations, each consisting of 50 epochs, and the number of samples in an epoch t is approximately $20(1.05)^t(S + A)K$.

Even though we did not theoretically analyze the VI-based method in this work, for the reasons mentioned in Section 3.3, we note that this method works well in practice for the settings considered in this study. We consider three methods: VI with leveraged anchors, where we use half of the experiences to estimate leverage scores and based on them sample the second half in a CUR-like fashion. Next, we consider VI with uniform anchors, where anchors are chosen uniformly at random without repetitions. And finally, we consider full-matrix VI, a standard VI approach without any matrix completion steps, where each entry of the matrix gets observed a certain number of times.

We see in Figure 4 that VI with leveraged anchors achieves the best performance measured in Frobenius and entrywise norm. On the other hand, VI with uniform-anchors does not recover specific entries with high values well (as seen in the right figure), but because there are not too many entries with high values, it achieves decent performance in the Frobenius norm. Finally, even though full-matrix VI can observe all entries of the matrix, it still lags behind VI with leveraged anchors. We also want to remind the reader that VI with leveraged anchors uses only half of the available samples for matrix recovery, while the other half is used for learning the leverage scores.

The algorithm used in the experimental section of [35] closely resembles our `LoRa-VI` algorithm when uniform anchors are applied. As a result, the numerical results from [35] can be reproduced within our framework, which offers a more general and flexible setting.

A.5 Low-rank Policy Iteration

Finally, we experimentally study performance of the proposed algorithm `LoRa-PI`.

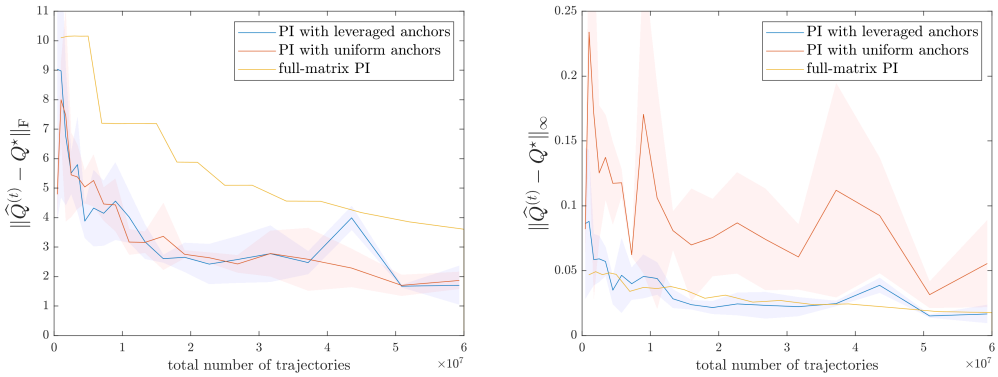


Figure 5: `LoRa-PI`: $Q^\star$ generated from low-rank r and P of rank $d = 4$, $S = A = 1000$, $\gamma = 0.1$, $\tau = 5$. We used $K = 10$ anchors, uniformly random initial policy, and noisy rewards with Gaussian noise $\sigma = 0.01$. Plots for PI with anchors are averaged over 3 simulations, while the one for full-matrix PI is simulated once. Each simulation consisted of 20 epochs, and the number of samples in an epoch t is approximately $10(1.15)^t(S + A)K$.

Similarly as in the previous subsection we study performance of three different methods using PI instead of VI this time, and the observed performance is similar to the one of VI-based methods. In contrast to the other methods, using leverage scores seems to ensure that Frobenius error behaves similarly to entrywise error, up to a scaling factor. This might be caused by uniform dispersion of the estimation error over the entries with large values for PI/VI with leveraged anchors.

The choice of $\gamma = 0.1$ is governed by an observation that this value of parameter γ ensures that the largest singular values of $Q^\star$ are scaling similarly. In other words, it is a heuristic for ensuring small κ needed for CUR-like methods. Furthermore, we believe that performance could be improved if a more tuned way of pseudoinversion is used. Namely, as $Q^{(t)}$ is effectively rank-deficient for many epochs, it is crucial to implement a stable way of calculating the pseudoinverse of $L\tilde{Q}(\mathcal{I}, \mathcal{J})R$, and make it dependent on the current epoch and the level of the estimation error.

Lastly, we believe that the performance of the proposed methods can be significantly improved (compared to full-matrix methods) for larger state-action spaces, as well as by implementing a more advanced way of distributing samples across epochs.

B Leverage Scores Estimation Analysis

In this section we provide the proof of Theorem 1. The proof relies on a tight instance dependent row-wise guarantee on the singular subspace recovery which is provided in Theorem 4 together with a proof. Throughout this section and for brevity, we use the notation

$$T_\tau = \frac{T}{\tau+1}, \qquad \text{and} \qquad \bar{\alpha} = \frac{r_{\max}}{1-\gamma}\frac{\sqrt{SA}}{\sigma_d(Q^\pi)}$$

in the entirety of this section to denote the number of sampled trajectories T_τ, and spikiness-related parameter $\bar{\alpha}$ (recall definition of spikiness α from Section 3.2). Furthermore, recall the truncated value matrix Q^π_τ defined in Section 4.1, and let us define its corresponding approximation error by $\Delta = Q^\pi_\tau - Q^\pi$.

B.1 Instance-dependent row-wise singular subspace recovery

Below, we present Theorem 4, which as highlighted before, is crucial in deriving Theorem 1.

Theorem 4. *If $T_\tau = \widetilde{\Omega}\left(\bar{\alpha}^2(S+A)\right)$ and $\|\Delta\|_{\text{op}} \le \sigma_d(Q^\pi)/32$, then we have that the event: for every $i \in [S]$, $j \in [A]$*

$$\|U_{i,:} - \widehat{U}_{i,:}O_{\widehat{U}}\|_2 = \widetilde{O}\left[\bar{\alpha}\left(\sqrt{\frac{d}{T_\tau}} + \kappa\|U_{i,:}\|_2\sqrt{\frac{S+A}{T_\tau}}\right) + \frac{\sqrt{S+A}\|\Delta\|_\infty}{\sigma_d(Q^\pi)} + \kappa\|U_{i,:}\|_2\frac{\|\Delta\|_{\text{op}}}{\sigma_d(Q^\pi)}\right],$$

$$\|W_{j,:} - \widehat{W}_{j,:}O_{\widehat{W}}\|_2 = \widetilde{O}\left[\bar{\alpha}\left(\sqrt{\frac{d}{T_\tau}} + \kappa\|W_{j,:}\|_2\sqrt{\frac{S+A}{T_\tau}}\right) + \frac{\sqrt{S+A}\|\Delta\|_\infty}{\sigma_d(Q^\pi)} + \kappa\|W_{j,:}\|_2\frac{\|\Delta\|_{\text{op}}}{\sigma_d(Q^\pi)}\right],$$

holds with probability at least $1-\delta$, where we define $O_{\widehat{U}} = \widehat{U}^\top U$ and $O_{\widehat{W}} = \widehat{W}^\top W$.

Corollary 1. *If $\|\Delta\|_\infty \le \min\left\{\frac{r_{\max}}{1-\gamma}\sqrt{\frac{d(S\wedge A)}{T_\tau}}, \frac{\sigma_d(Q^\pi)}{32\sqrt{SA}}\right\}$ and $T_\tau = \widetilde{\Omega}\left(\bar{\alpha}^2(S+A)\right)$, then w.h.p:*

$$\|U_{i,:} - \widehat{U}_{i,:}(\widehat{U}^\top U)\|_2 = \widetilde{O}\left[\bar{\alpha}\left(\sqrt{\frac{d}{T_\tau}} + \sqrt{\frac{S+A}{T_\tau}}\kappa\|U_{i,:}\|_2\right)\right].$$

An analogous inequality holds for $\|W_{i,:} - \widehat{W}_{i,:}(\widehat{W}^\top W)\|_2$.

It is a simple algebraic exercise to show that $\epsilon = \|\Delta\|_\infty$ from (1) satisfies condition of the corollary above in given regime of T_τ.

B.2 Proof of Theorem 1

Proof. First we consider those states with $\|U_{s,:}\|_2^2 > \frac{d}{4S}$. From Corollary 1 we obtain that for these states and large enough T_τ:

$$\|U_{s,:} - \widehat{U}_{s,:}(\widehat{U}^\top U)\|_2 \le c_1\bar{\alpha}\left(\sqrt{\frac{d}{T_\tau}} + \sqrt{\frac{S+A}{T_\tau}}\kappa\|U_{s,:}\|_2\right)\log^{3/2}\left(\frac{T_\tau(S+A)}{\delta}\right)$$

w.h.p. and for some universal constant $c_1 > 0$. Since $\|U_{s,:}\|_2 > \sqrt{\frac{d}{4S}}$ this implies that $\|U_{s,:}\|_2 > \sqrt{\frac{d}{4(S+A)\kappa^2}}$, we can simplify last inequality as follows:

$$\|U_{s,:} - \widehat{U}_{s,:}(\widehat{U}^\top U)\|_2 \le 2c_1\bar{\alpha}\sqrt{\frac{S+A}{T_\tau}}\kappa\|U_{s,:}\|_2\log^{3/2}\left(\frac{T_\tau(S+A)}{\delta}\right)$$

Next, for $T_\tau \ge 50c_1^2\bar{\alpha}^2(S+A)\kappa^2\log^3\left(\frac{T_\tau(S+A)}{\delta}\right)$, we have: $\|U_{s,:} - \widehat{U}_{s,:}(\widehat{U}^\top U)\|_2 \le (1 - \frac{1}{\sqrt{2}})\|U_{s,:}\|_2$. Finally, using reverse triangle inequality we obtain:

$$\|\widehat{U}_{s,:}\|_2^2 \ge (\|U_{s,:}\|_2 - \|U_{s,:} - \widehat{U}_{s,:}(\widehat{U}^\top U)\|_2)^2 \ge \frac{1}{2}\|U_{s,:}\|_2^2$$

and thus:

$$\tilde{\ell}_s = \|\widehat{U}_{s,:}\|_2^2 \vee \frac{d}{S} \geq \|\widehat{U}_{s,:}\|_2^2 \geq \frac{1}{2}\|U_{s,:}\|_2^2$$

Now we consider states with $\|U_{s,:}\|_2^2 \leq \frac{d}{4S}$. Again, by means of Corollary 1 we get w.h.p:

$$\|U_{s,:} - \widehat{U}_{s,:}(\widehat{U}^\top U)\|_2 \leq 2c_1\bar{\alpha}\kappa\sqrt{\frac{d}{T_\tau}}\sqrt{\frac{S+A}{S}}\log^{3/2}\left(\frac{T_\tau(S+A)}{\delta}\right) \leq \sqrt{\frac{d}{4S}}$$

for $T_\tau \geq 16c_1^2\bar{\alpha}^2\kappa^2(S+A)\log^3\left(\frac{T_\tau(S+A)}{\delta}\right)$. Thus we obtain that for all s with $\|U_{s,:}\|_2^2 \leq \frac{d}{4S}$, it also holds:

$$\|\widehat{U}_{s,:}\|_2 \leq \|U_{s,:}\|_2 + \|U_{s,:} - \widehat{U}_{s,:}(\widehat{U}^\top U)\|_2 \leq \sqrt{\frac{d}{S}}$$

Since by definition $\tilde{\ell}_s \geq \frac{d}{S}$, we obtain that $\tilde{\ell}_s \geq \|U_{s,:}\|_2^2$ for states with $\|U_{s,:}\|_2^2 \leq \frac{d}{4S}$.

Finally, we show similar inequalities hold for leverage scores ℓ and $\hat{\ell}$. Namely, we have:

$$\begin{aligned}\hat{\ell}_s = \frac{\tilde{\ell}_s}{\|\tilde{\ell}\|_1} &\geq \frac{\tilde{\ell}_s}{\sum_{i:\|\widehat{U}_{i,:}\|_2^2 > \frac{d}{S}}\|\widehat{U}_{i,:}\|_2^2 + \sum_{j:\|\widehat{U}_{j,:}\|_2^2 \leq \frac{d}{S}}\frac{d}{S}} \\ &\geq \frac{\tilde{\ell}_s}{\sum_{i=1}^S \|\widehat{U}_{i,:}\|_2^2 + S\frac{d}{S}} = \frac{\tilde{\ell}_s}{2d} \geq \frac{\|U_{s,:}\|_2^2}{4d} = \frac{1}{4}\ell_s\end{aligned}$$

where we used first part of the proof for the final inequality.

□

B.3 Proof of Theorem 4

Proof is based on leave-one-out method used for proving entry.wise guarantees for singular vectors of SVD estimates. We refer the interested reader to [8] for a comprehensive survey about the method. Here, we repeat the main arguments of the proof and improve the analysis in the following two ways:

(i) We keep track of approximation error during the whole proof in order to be able to apply it to approximately low rank matrix Q_τ^π;

(ii) Instead of showing guarantees in $\|\cdot\|_{2\to\infty}$ norm, we prove row/column specific guarantees. Indeed, note that our guarantees do not depend explicitly on the incoherence parameter μ but instead the guarantee specific to the row vector $U_{i,:}$ of the singular subspace U, depends only on its own incoherence parameter, i.e., $\|U_{i,:}\|_2$. This enables us to do leverage score analysis and obtain Theorem 1.

Leave-one-out method is applied to symmetric matrices, so we first redefine our matrices in this context. For a matrix $Q^\pi \in \mathbb{R}^{S\times A}$ with SVD $Q^\pi = U\Sigma W^\top$ define symmetrizated matrix M_d as follows:

$$M_d = \begin{bmatrix} 0 & Q^\pi \\ (Q^\pi)^\top & 0 \end{bmatrix} = \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix} U & U \\ W & -W \end{bmatrix}}_{\mathbf{U}}\underbrace{\begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}}_{\mathbf{\Sigma}}\underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix} U & U \\ W & -W \end{bmatrix}^\top}_{\mathbf{U}^\top}$$

and similarly define symmetrized matrix M from Q_τ^π, $\mathbf{\Delta}$ from Δ, $\widehat{M}_d$ from $\widehat{Q}^\pi$, and $\mathbf{E}$ from $E = \widetilde{Q}_\tau^\pi - Q_\tau^\pi$. Note that, using this notation, we have $M = M_d + \mathbf{\Delta}$, with rank $d' = 2d$ matrix M_d and $\|\mathbf{\Delta}\|_\infty = \|\Delta\|_\infty, \|\mathbf{\Delta}\|_{\text{op}} = \|\Delta\|_{\text{op}}$. Assume observation matrix is given by $\widetilde{M} = M + \mathbf{E} = M_d + \mathbf{\Delta} + \mathbf{E}$ and note that $\widehat{M}_d = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{U}}^\top$ and $\widehat{M}_d = \widetilde{M}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top$. Thus, in order to prove lemma, it is sufficient to show:

$$\begin{aligned}\|\mathbf{U}_{i,:} - \widehat{\mathbf{U}}_{i,:}(\widehat{\mathbf{U}}^\top\mathbf{U})\|_2 \lesssim \left(\sqrt{d'} + \kappa\|\mathbf{U}_{i,:}\|_2\sqrt{S+A}\right)\frac{r_{\max}}{1-\gamma}\frac{\sqrt{SA}}{\sqrt{T_\tau}\sigma_{d'}(M)}\log^{3/2}\left(\frac{T_\tau(S+A)}{\delta}\right) \\ + \frac{\sqrt{S+A}\|\mathbf{\Delta}\|_\infty}{\sigma_{d'}(M)} + \kappa\|\mathbf{U}_{i,:}\|_2\frac{\|\mathbf{\Delta}\|_{\text{op}}}{\sigma_{d'}(M)}\end{aligned}$$

For now, let us assume that $\|\mathbf{E}\|_{\text{op}} \leq \sigma_{d'}(M)/32$, and prove that such inequality holds for T_τ large enough (consequence of Proposition 2).

Define $\widetilde{M}_d = M_d + \mathbf{E}$. Fix any $i \in [S+A]$ and for any matrix $A \in \mathbb{R}^{(S+A)\times n}$ define seminorm: $\|A\|_{2,i} = \|A_{i,:}\|_2$. Now using that $\mathbf{U} = M_d\mathbf{U}\mathbf{\Sigma}^{-1} = (\widetilde{M}_d - \mathbf{E})\mathbf{U}\mathbf{\Sigma}^{-1}$, we have:

$$\|\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{2,i} \leq \frac{\|\widetilde{M}_d\mathbf{U} - \widehat{\mathbf{U}}(\widehat{\mathbf{U}}^\top\mathbf{U})\mathbf{\Sigma}\|_{2,i}}{\sigma_{d'}(M)} + \frac{\|\mathbf{E}\mathbf{U}\|_{2,i}}{\sigma_{d'}(M)}$$

To bound the numerator of the first term, note that $\widehat{\mathbf{U}}(\widehat{\mathbf{U}}^\top\mathbf{U})\mathbf{\Sigma} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top M_d\mathbf{U} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top(\widetilde{M} - \mathbf{\Delta} - \mathbf{E})\mathbf{U}$. Since $\widehat{\mathbf{U}}^\top\widetilde{M} = \widehat{\mathbf{U}}^\top\widehat{M}_d = \widehat{\mathbf{\Sigma}}\widehat{\mathbf{U}}^\top$, we get

$$\begin{aligned}
\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{\Sigma} &= \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{U}}^\top\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top(\mathbf{E}+\mathbf{\Delta})\mathbf{U} \\
&= \widetilde{M}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top(\mathbf{E}+\mathbf{\Delta})\mathbf{U} \\
&= \widetilde{M}_d\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} + \mathbf{\Delta}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top(\mathbf{E}+\mathbf{\Delta})\mathbf{U} \\
&= \widetilde{M}_d\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} + \mathbf{\Delta}(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} - \mathbf{U}) + \mathbf{\Delta}\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top(\mathbf{E}+\mathbf{\Delta})\mathbf{U}
\end{aligned}$$

Consequently, for $(\star) := \|\widetilde{M}_d\mathbf{U} - \widehat{\mathbf{U}}(\widehat{\mathbf{U}}^\top\mathbf{U})\mathbf{\Sigma}\|_{2,i}$ we have:

$$\begin{aligned}
(\star) \leq \|M_d(\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i} + \|\mathbf{E}(\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i} + \|\mathbf{\Delta}(\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i} \\
+ \|\mathbf{\Delta}\mathbf{U}\|_{2,i} + \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top(\mathbf{E}+\mathbf{\Delta})\mathbf{U}\|_{2,i}
\end{aligned}$$

Throughout the proof we use Davis-Kahan's inequality (Corollary 2.8 in [8]): if $\|\mathbf{E}\|_{\text{op}} < (1 - 1/\sqrt{2})\sigma_{d'}(M)$, then:

$$\|\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{\text{op}} \leq \frac{2\|\mathbf{E}\|_{\text{op}}}{\sigma_{d'}(M)}$$

Now, similarly to (39) in [37] and using that $\|AB\|_{2,i} \leq \|A\|_{2,i}\|B\|_{\text{op}}$ we obtain:

$$\begin{aligned}
\|M_d(\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i} &\leq 4\|\mathbf{U}\|_{2,i}\|\mathbf{\Sigma}\|_{\text{op}}\frac{\|\mathbf{E}\|_{\text{op}}^2}{\sigma_{d'}^2(M)} \\
\|\mathbf{\Delta}(\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i} &\leq 2\|\mathbf{\Delta}\|_{2,i}\frac{\|\mathbf{E}\|_{\text{op}}}{\sigma_{d'}(M)} \\
\|\mathbf{\Delta}\mathbf{U}\|_{2,i} &\leq \|\mathbf{\Delta}\|_{2,i}\|\mathbf{U}\|_{\text{op}} \leq \|\mathbf{\Delta}\|_{2,i}
\end{aligned}$$

Using these inequalities together with Lemma 3 and $\frac{\|\mathbf{E}\|_{\text{op}} + \|\mathbf{\Delta}\|_{\text{op}}}{\sigma_{d'}(M)} \leq 1/16$ we obtain:

$$(\star) \leq 2\|\mathbf{\Delta}\|_{2,i} + \frac{9}{2}\|\mathbf{U}\|_{2,i}\frac{\|\mathbf{\Sigma}\|_{\text{op}}}{\sigma_{d'}(M)}(\|\mathbf{E}\|_{\text{op}} + \|\mathbf{\Delta}\|_{\text{op}}) + 2\|\mathbf{E}(\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i} + \|\mathbf{E}\mathbf{U}\|_{2,i}$$

which in the end gives

$$\begin{aligned}
\|\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{2,i} \leq \frac{2}{\sigma_{d'}(M)}\Big(5\|\mathbf{U}\|_{2,i}\frac{\|\mathbf{\Sigma}\|_{\text{op}}}{\sigma_{d'}(M)}(\|\mathbf{E}\|_{\text{op}} + \|\mathbf{\Delta}\|_{\text{op}}) + \|\mathbf{E}\mathbf{U}\|_{2,i} \\
+ \|\mathbf{E}(\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i} + \|\mathbf{\Delta}\|_{2,i}\Big) \qquad (9)
\end{aligned}$$

Leave-one-out decomposition: Define matrix $\widetilde{M}^{(i)}$ as follows:

$$\widetilde{M}^{(i)}_{j,k} = \begin{cases} \widetilde{M}_{j,k}, & \text{if } j \neq i \text{ or } k \neq i \\ M_{j,k}, & \text{otherwise} \end{cases}$$

and let $\widehat{\mathbf{U}}^{(i)}$ denote matrix of d' dominant singular vectors of $\widetilde{M}^{(i)}$. Then, by triangle inequality we can write:

$$\|\mathbf{E}(\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i} \leq \|\mathbf{E}(\mathbf{U} - \widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top\mathbf{U})\|_{2,i} + \|\mathbf{E}\|_{\text{op}}\|\widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top\mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{\text{F}}$$

We bound the last term using Lemma 4 to obtain:

$$\|\mathbf{E}(\mathbf{U}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i} \le 2\|\mathbf{E}(\mathbf{U}-\widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top\mathbf{U})\|_{2,i} + 6\frac{\|\mathbf{E}\|_{\mathrm{op}}}{\sigma_{d'}(M)}\Big(\|\mathbf{E}\mathbf{U}\|_{2,i} + 2\|\mathbf{E}\|_{\mathrm{op}}(\|\mathbf{U}\|_{2,i} + \|\mathbf{U}-\widehat{\mathbf{U}}(\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i})\Big) \tag{10}$$

Substituting (10) into (9) we obtain:

$$\|\mathbf{U}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{2,i} \le \frac{12}{\sigma_{d'}(M)}\Big(\|\mathbf{U}\|_{2,i}\frac{\|\mathbf{\Sigma}\|_{\mathrm{op}}}{\sigma_{d'}(M)}(\|\mathbf{E}\|_{\mathrm{op}}+\|\mathbf{\Delta}\|_{\mathrm{op}}) + \|\mathbf{E}\mathbf{U}\|_{2,i} + \|\mathbf{E}(\mathbf{U}-\widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top\mathbf{U})\|_{2,i} + \|\mathbf{\Delta}\|_{2,i}\Big)$$

Then we apply Proposition 2 on $\|\mathbf{E}\mathbf{U}\|_{2,i}$ and $\|\mathbf{E}(\mathbf{U}-\widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top\mathbf{U})\|_{2,i}$. We use that $\|\mathbf{U}\|_{\mathrm{F}} \le \sqrt{d'}$ and $\|\mathbf{U}-\widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top\mathbf{U}\|_{\mathrm{F}} \le 2\sqrt{d'}$. Finally, we use that $\|\mathbf{\Delta}\|_{2,i} \le \sqrt{S+A}\|\Delta\|_\infty$, the fact that

$$\sigma_{d'}(M) \ge \sigma_{d'}(M_d) - \|\mathbf{\Delta}\|_{\mathrm{op}} \ge \sigma_{d'}(M_d)(1-1/32)$$

and thus $\frac{\|\mathbf{\Sigma}\|_{\mathrm{op}}}{\sigma_{d'}(M)} \le 2\frac{\|M_d\|_{\mathrm{op}}}{\sigma_{d'}(M_d)} \le 2\kappa$.

B.4 Rank estimation guarantee

Recall from Section 4.2 and Proposition 1 that, given the singular values $(\hat{\sigma}_i)_i$ of our estimates $\widetilde{Q}^\pi_\tau$, we estimate effective rank as follows: $\hat{d} = \sum_{i=1}^{S\wedge A} \mathbb{1}\{\hat{\sigma}_i \ge \beta\}$ with

$$\beta = \sqrt{\frac{r_{\max}^2 SA(S+A)}{(1-\gamma)^3 T}\log^4\left(\frac{(S+A)T}{(1-\gamma)\delta}\right)} + \frac{r_{\max}\sqrt{SA}}{T}$$

Here we repeat first part of the Proposition 1 and prove it:

Lemma 2. *If T satisfies (5), then estimated rank $\hat{d}$ satisfies $\hat{d} = d_\pi$ with probability at least $1-\delta$.*

Proof. By our assumptions $\sigma_i(Q^\pi) > 0$ only for $i \in [d_\pi]$, and thus $\forall i > d_\pi:\ \hat{\sigma}_i \le \|\widetilde{Q}^\pi_\tau - Q^\pi\|_{\mathrm{op}}$ and $\forall i \le d_\pi:\ \hat{\sigma}_i \ge \sigma_{d_\pi}(Q^\pi) - \|\widetilde{Q}^\pi_\tau - Q^\pi\|_{\mathrm{op}}$. Recall that $E = \widetilde{Q}^\pi_\tau - Q^\pi_\tau$ and $\Delta = Q^\pi_\tau - Q^\pi$, and thus:

$$\|\widetilde{Q}^\pi_\tau - Q^\pi\|_{\mathrm{op}} \le \|E\|_{\mathrm{op}} + \|\Delta\|_{\mathrm{op}}$$

We bound the second term by $\|\Delta\|_{\mathrm{op}} \le \sqrt{SA}\|\Delta\|_\infty$ and use that $\|\Delta\|_\infty \le \frac{r_{\max}}{T}$ from Lemma 1. The first term is upper bounded by Proposition 2. Combining the two, we obtain that $\|\widetilde{Q}^\pi_\tau - Q^\pi\|_{\mathrm{op}} \le \beta$ with high probability. Thus, for $2\beta \le \sigma_{d_\pi}(Q^\pi)$ we are guaranteed to recover the true rank d_π, since then $\forall i \in [d_\pi]$:

$$\hat{\sigma}_i \ge \sigma_{d_\pi}(Q^\pi) - \|\widetilde{Q}^\pi_\tau - Q^\pi\|_{\mathrm{op}} \ge 2\beta - \beta \ge \beta$$

It is straightforward to check that, if

$$T = \Omega\left(\frac{r_{\max}^2 SA(S+A)}{(1-\gamma)^3\sigma_d^2(Q^\pi)}\log^4\left(\frac{(S+A)T}{(1-\gamma)\delta}\right)\right)$$

then first term in definition of β is $\le \sigma_{d_\pi}(Q^\pi)/4$, and similar conclusion hold for the second term after noting that $\frac{r_{\max}\sqrt{SA}}{\sigma_{d_\pi}(Q^\pi)}(S+A) \ge 1$. □

B.5 Technical lemmas from the proof of Theorem 4

In this section we shortly present concentration results used in the proof of Theorem 4. We refer reader to Section F.2 for discussion about equivalent noise model and Poisson approximation. Concentration inequalities proposed are fairly standard (see for example [8]), but as discussed in [37] because of the sampling pattern, entries of the matrix $\mathbf{E}$ are slightly dependent. A way to deal with these dependencies has been discussed in [37], and we refer to the results from that paper here directly.

Proposition 2. *Let B be a $(S+A)\times 2d$ matrix independent of $\mathbf{E}$. Then, we have for all $\delta\in(0,1)$, for all $T_\tau \gtrsim (S+A)\log^3((S+A)/\delta)$, both events:*

$$\|\mathbf{E}\|_{\mathrm{op}} \lesssim \frac{r_{\max}}{1-\gamma}\sqrt{\frac{SA}{T_\tau}}\sqrt{S+A}\log^{3/2}\left(\frac{T_\tau(S+A)}{\delta}\right)$$

$$\forall i\in[S+A]: \quad \|\mathbf{E}_{i,:}B\|_{\mathrm{op}} \lesssim \frac{r_{\max}}{1-\gamma}\|B\|_{\mathrm{F}}\sqrt{\frac{SA}{T_\tau}}\log^{3/2}\left(\frac{T_\tau(S+A)}{\delta}\right)$$

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

Proof. Follows straightforwardly from proofs of Propositions 26 and 27 in [37] and using noise equivalent model from Section F.2. Note that we keep the variance term upper bounded by $\|A\|_{\mathrm{F}}^2$ in the proof of Proposition 27 as in [37], and make use of inequality $\|B\|_{2\to\infty}\le\|B\|_{\mathrm{F}}$ to obtain dependence on $\|B\|_{\mathrm{F}}$ in the second inequality. □

Lemma 3. *If $\|\mathbf{E}\|\le\sigma_{d'}(M)/32$, then for every i:*

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top(\mathbf{E}+\boldsymbol{\Delta})\mathbf{U}\|_{2,i} \le 4\frac{\|\mathbf{E}\|_{\mathrm{op}}+\|\boldsymbol{\Delta}\|_{\mathrm{op}}}{\sigma_{d'}(M)}\Big(\|\mathbf{U}\|_{2,i}\|\boldsymbol{\Sigma}\|_{\mathrm{op}}+\|\boldsymbol{\Delta}\|_{2,i} + \|\mathbf{E}\mathbf{U}\|_{2,i}+\|\widetilde{M}(\mathbf{U}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i}\Big)$$

Proof. Under condition $\|\mathbf{E}\|\le\sigma_{d'}(M)/32$ we have $\sigma_{d'}(\widetilde{M})\ge\sigma_{d'}(M)-\|\mathbf{E}\|\ge\sigma_{d'}(M)/2$. Thus, we have:

$$\begin{aligned}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top(\mathbf{E}+\boldsymbol{\Delta})\mathbf{U}\|_{2,i} &= \|\widehat{M}_d\widehat{\mathbf{U}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{U}}^\top(\mathbf{E}+\boldsymbol{\Delta})\mathbf{U}\|_{2,i} = \|\widetilde{M}\widehat{\mathbf{U}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{U}}^\top(\mathbf{E}+\boldsymbol{\Delta})\mathbf{U}\|_{2,i}\\ &\le \|\widetilde{M}\widehat{\mathbf{U}}\|_{2,i}\|\widehat{\boldsymbol{\Sigma}}^{-1}\|_{\mathrm{op}}\|\widehat{\mathbf{U}}^\top\|_{\mathrm{op}}\|\mathbf{E}+\boldsymbol{\Delta}\|_{\mathrm{op}}\|\mathbf{U}\|_{\mathrm{op}} \le \|\widetilde{M}\widehat{\mathbf{U}}\|_{2,i}\frac{\|\mathbf{E}+\boldsymbol{\Delta}\|_{\mathrm{op}}}{\sigma_{d'}(\widetilde{M})}\\ &\le 2\frac{\|\mathbf{E}\|_{\mathrm{op}}+\|\boldsymbol{\Delta}\|_{\mathrm{op}}}{\sigma_{d'}(M)}\|\widetilde{M}\widehat{\mathbf{U}}\|_{2,i} \qquad (11)\end{aligned}$$

Using Davis-Kahan's inequality [8] we have:

$$\begin{aligned}\|\widetilde{M}\widehat{\mathbf{U}}\|_{2,i} &= \|\widetilde{M}\widehat{\mathbf{U}}\mathrm{sgn}(\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i}\\ &\le \|\widetilde{M}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{2,i}+\|\widetilde{M}\widehat{\mathbf{U}}\|_{2,i}\|\mathrm{sgn}(\widehat{\mathbf{U}}^\top\mathbf{U})-\widehat{\mathbf{U}}^\top\mathbf{U}\|_{\mathrm{op}}\\ &\le \|\widetilde{M}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{2,i}+16\|\widetilde{M}\widehat{\mathbf{U}}\|_{2,i}\frac{\|\mathbf{E}\|^2}{\sigma_{d'}(M)^2}\\ &\le \|\widetilde{M}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{2,i}+\frac{\|\widetilde{M}\widehat{\mathbf{U}}\|_{2,i}}{2}\end{aligned}$$

implying that $\|\widetilde{M}\widehat{\mathbf{U}}\|_{2,i}\le 2\|\widetilde{M}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{2,i}$. Furthermore, we have:

$$\begin{aligned}\|\widetilde{M}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{2,i} &\le \|\widetilde{M}\mathbf{U}\|_{2,i}+\|\widetilde{M}(\mathbf{U}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i}\\ &\le \|\mathbf{U}\|_{2,i}\|\boldsymbol{\Sigma}\|_{\mathrm{op}}+\|(\mathbf{E}+\boldsymbol{\Delta})\mathbf{U}\|_{2,i}+\|\widetilde{M}(\mathbf{U}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i}\end{aligned}$$

where we have used that $M\mathbf{U}=\mathbf{U}\boldsymbol{\Sigma}$. Substituting this expression back into (11) finishes the proof. □

Lemma 4. *Under assumptions $\|\boldsymbol{\Delta}\|_{\mathrm{op}}\le\sigma_{d'}(M)/32$ and $\|\mathbf{E}\|_{\mathrm{op}}\le\sigma_{d'}(M)/32$, we have with high probability for every i:*

$$\|\widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top\mathbf{U}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{\mathrm{F}} \le \frac{6}{\sigma_{d'}(M)}\Big(\|\mathbf{E}\mathbf{U}\|_{2,i}+\|\mathbf{E}(\mathbf{U}-\widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top\mathbf{U})\|_{2,i} + 2\|\mathbf{E}\|_{\mathrm{op}}(\|\mathbf{U}\|_{2,i}+\|\mathbf{U}-\widehat{\mathbf{U}}(\widehat{\mathbf{U}}^\top\mathbf{U})\|_{2,i})\Big)$$

Proof. Proof follows similar steps as Step 2.2 in the proof of Theorem 4.2 in [8], but we repeat it here for the sake of completeness and focus on differences in the proof caused by $\mathbf{\Delta}$ matrix. First, we use that $\mathbf{U}$ is an orthogonal matrix ($\|\mathbf{U}\|_{\text{op}} = 1$) and Davis-Kahan's inequality to obtain:

$$\|\widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top \mathbf{U} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{U}\|_{\text{F}} \leq \|\widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\|_{\text{F}} \leq 2\frac{\|(\widetilde{M} - \widetilde{M}^{(i)})\widehat{\mathbf{U}}^{(i)}\|_{\text{F}}}{\sigma_{d'}(\widetilde{M}^{(i)}) - \sigma_{d'+1}(\widetilde{M}^{(i)})}$$

Note that under the assumption $\|\mathbf{\Delta}\|_{\text{op}} \leq \sigma_{d'}(M)$ analysis of singular values stays the same as in [8], since, for example:

$$\sigma_{d'}(\widetilde{M}^{(i)}) \geq \sigma_{d'}(M) - \|\mathbf{E}^{(i)}\|_{\text{op}} \geq \sigma_{d'}(M)\left(1 - \frac{1}{32}\right)$$
$$\sigma_{d'+1}(\widetilde{M}^{(i)}) \leq \sigma_{d'+1}(M) + \|\mathbf{E}^{(i)}\|_{\text{op}} \leq \|\mathbf{\Delta}\|_{\text{op}} + \|\mathbf{E}\|_{\text{op}} \leq \sigma_{d'}(M)/16$$

Thus, we can lower bound denominator in the inequality above by $\sigma_{d'}(M)/2$. Now, term in the numerator can be written as:

$$(\widetilde{M} - \widetilde{M}^{(i)})\widehat{\mathbf{U}}^{(i)} = \mathbf{E}_{i,:}\widehat{\mathbf{U}}^{(i)} + (\mathbf{E}_{:,i} - \mathbf{E}_{i,i}e_i)\widehat{\mathbf{U}}^{(i)}_{i,:}$$

and bounded in the same way as in [8]:

$$\begin{aligned}\|(\widetilde{M} - \widetilde{M}^{(i)})\widehat{\mathbf{U}}^{(i)}\|_{\text{F}} &\leq \|\mathbf{E}\widehat{\mathbf{U}}^{(i)}\|_{2,i} + \|\mathbf{E}_{:,i} - \mathbf{E}_{i,i}e_i\|_2\|\widehat{\mathbf{U}}^{(i)}\|_{2,i} \\ &\leq \|\mathbf{E}\widehat{\mathbf{U}}^{(i)}\|_{2,i} + 2\|\mathbf{E}\|_{\text{op}}\|\widehat{\mathbf{U}}^{(i)}((\widehat{\mathbf{U}}^{(i)})^\top \mathbf{U})\|_{2,i}\end{aligned}$$

where we used that $\|((\widehat{\mathbf{U}}^{(i)})^\top \mathbf{U})^{-1}\|_2 \leq 2$ under our assumptions. Finally, we obtain:

$$\begin{aligned}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{U} - \widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top \mathbf{U}\|_{\text{F}} \leq \frac{4}{\sigma_{d'}(M)}(&\|\mathbf{E}\widehat{\mathbf{U}}^{(i)}\|_{2,i} + 2\|\mathbf{E}\|_{\text{op}}\|\widehat{\mathbf{U}}(\widehat{\mathbf{U}}^\top \mathbf{U})\|_{2,i} \\ &+ 2\|\mathbf{E}\|_{\text{op}}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{U} - \widehat{\mathbf{U}}^{(i)}(\widehat{\mathbf{U}}^{(i)})^\top \mathbf{U}\|_{\text{F}})\end{aligned}$$

and under condition $\|\mathbf{E}\|_{\text{op}} \leq \sigma_{d'}(M)/32$ we obtain result claimed in the lemma. □

B.6 Nuclear norm minimization for leverage scores estimation

The authors of [34] leveraged the guarantees for nuclear norm minimization from [9] to learn Q matrices. However, several factors make the application of nuclear norm minimization theoretically challenging in our context:

- **Approximate low-rank structure.** As our algorithm is based on policy iteration, our estimates Q^π_τ are only approximately low-rank. The result from [9] rely on non-convex optimization, leaving it unclear how this approximation error affects the final guarantees of the algorithm. In contrast, a more straightforward analysis using singular value decomposition allows us to explicitly express our bounds in terms of the approximation error.
- **Coherence-free subspace recovery.** In our subspace recovery result (Theorem 4), we can bound the subspace error $\|U_{i,:} - \widehat{U}_{i,:}O_{\widehat{U}}\|_2$ in relation to $\|U_{i,:}\|_2$. It is uncertain whether current guarantees for nuclear norm minimization can achieve a similar outcome, which might instead depend on $\max_{i\in[S]}\|U_{i,:}\|_2$. We believe this would introduce dependency on the incoherence constant into the sample complexity of our algorithm.

C Leveraged Matrix Estimation Analysis

In this appendix, we provide the proof of Theorem 2 corresponding to sample complexity guarantee enjoyed by LME. First, we provide the pseudo-code of LME:

Algorithm 2: Leverage Matrix Estimation (LME)

Input: Deterministic policy π, sampling budget T
Set $T_1 \leftarrow T/2, T_2 \leftarrow T/2$
Set $\epsilon = \frac{r_{\max}}{T}$, $\tau \leftarrow \frac{1}{(1-\gamma)} \log\left(\frac{T}{1-\gamma}\right)$ as in (1)
(Phase 1). Leverage Scores Estimation:
(Phase 1a.) Data Collection & Emprirical Truncated Value matrix.
Sample uniformly at random $T_1/(\tau+1)$ trajectories of length $\tau+1$ using policy π and gather them in $\mathcal{D}$
Use the collected dataset $\mathcal{D}$, to construct $\widetilde{Q}^{\pi}_{\tau}$ as in (2)
(Phase 1b.) Singular Subspace Recovery
Set the threshold β as in (4)
Compute the SVD of $\widetilde{Q}^{\pi}_{\tau}$ and threshold with β as described in (3) to obtain $\widehat{d}$, $\widehat{U}$, $\widehat{W}$ and $\widehat{Q}^{\pi}$
(Phase 1c.) Leverage Scores.
Set the left (resp. right) leverage scores $\hat{\ell}$ (resp. $\hat{\rho}$) as described in (6).
(Phase 2.) CUR-based Matrix Estimation with Leverage.
(Phase 2a.) Data Collection with Leverage & Empirical Truncated Value Matrix:
Set $K \leftarrow 64\hat{d}\log(64\hat{d}/\delta)$
Sample K rows (resp. columns) $\mathcal{I} \subset \mathcal{S}$ (resp. $\mathcal{J} \subset \mathcal{A}$) without replacement according to the leverage scores $\hat{\ell}$ (resp. $\hat{\rho}$)
Set $N_1 \leftarrow \frac{T_2}{2(\tau+1)K^2}$, $N_2 \leftarrow \frac{T_2}{2(\tau+1)(K(S+A)-2K^2)}$
For all $(s,a) \in \Omega_{\square}$ (resp. $(s,a) \in \Omega_{+}$) sample N_1 (resp. N_2) trajectories of length $\tau+1$ using policy π and construct the empirical estimate $\widetilde{Q}^{\pi}_{\tau}(s,a)$ based on these trajectories
(Phase 2b. CUR-based Matrix estimation)
Set the matrices L and R as in (7)
Construct $\widehat{Q}^{\pi}$ using a CUR-based approach as in (8)
Output: $\widehat{Q}^{\pi}$.

C.1 Proof of Theorem 2

Before showing the proof of Theorem 2 we present two intermediate results used in the proof. As an immediate consequence of Hoeffding's inequality we have:

Lemma 5. *With probability at least $1-\delta$ we have $\forall (s,a) \in (\mathcal{I}\times\mathcal{A}) \cup (\mathcal{S}\times\mathcal{J})$:*

$$|\widetilde{Q}^{\pi}_{\tau}(s,a) - Q^{\pi}(s,a)| \leq \frac{r_{\max}}{1-\gamma}\sqrt{\frac{2}{N}\log\left(\frac{4K(S+A)}{\delta}\right)} + \|Q^{\pi}_{\tau} - Q^{\pi}\|_{\infty}$$

where $N = N_1$ if $(s,a) \in \Omega_{\square}$ or $N = N_2$ if $(s,a) \in \Omega_{+}$.

Theorem 5. *Let $\mathcal{I}$ and $\mathcal{J}$ be such that $|\mathcal{I}|, |\mathcal{J}| = K$, and $Q^{\pi}(\mathcal{I},\mathcal{J})$ has rank d. Assume that for some $\varepsilon_{\square}, \varepsilon_{+} > 0$:*

a) $\forall (s,a) \in \mathcal{I}\times\mathcal{J} : |\widetilde{Q}^{\pi}_{\tau}(s,a) - Q^{\pi}(s,a)| \leq \varepsilon_{\square}$, *and*

b) $\forall (s,a) \in (\mathcal{I}\times\mathcal{A}) \cup (\mathcal{S}\times\mathcal{J}) \setminus (\mathcal{I}\times\mathcal{J}) : |\widetilde{Q}^{\pi}_{\tau}(s,a) - Q^{\pi}(s,a)| \leq \varepsilon_{+}$.

If $\varepsilon_{\square} \leq \frac{1}{8c_{\mathcal{I}}c_{\mathcal{J}}}\frac{\sigma_d(Q^{\pi})}{\sqrt{SA}}\log^{-2}\left(\frac{S+A}{\delta}\right)$, $\varepsilon_{+} \leq \|Q^{\pi}\|_{\infty}$ and $K \geq 64d\log(64d/\delta)$, then with probability $\geq 1-\delta$:

$$\|\widehat{Q}^{\pi} - Q^{\pi}\|_{\infty} \lesssim \|Q^{\pi}\|_{\infty}\varepsilon_{+}\frac{\sqrt{SA}}{\sigma_d(Q^{\pi})}\log^2\left(\frac{S+A}{\delta}\right) + \|Q^{\pi}\|^2_{\infty}\varepsilon_{\square}\frac{SA}{\sigma^2_d(Q^{\pi})}\log^4\left(\frac{S+A}{\delta}\right)$$

Proof of Theorem 5 is deferred to C.2.

Proof of Theorem 2. First, by Theorem 1 we require at least

$$T \gtrsim \frac{r_{\max}^2}{(1-\gamma)^3\|Q^\pi\|_\infty^2}\kappa^2(S+A)\frac{\|Q^\pi\|_\infty^2 SA}{\sigma_d^2(Q^\pi)}\log^4\left(\frac{(S+A)T}{(1-\gamma)\delta}\right) \tag{12}$$

number of samples to recover leverage scores of Q^π. During the whole proof of Theorem 2 we condition on the event where Theorem 1 holds.

Recall that we use $N_1 = \frac{T}{4(\tau+1)K^2}$, $N_2 = \frac{T}{4(\tau+1)(K(S+A)-2K^2)}$ and define the following quantities:

$$\varepsilon_\square = \frac{r_{\max}}{1-\gamma}\sqrt{\frac{8}{N_1}\log\left(\frac{4K(S+A)}{\delta}\right)}, \qquad \varepsilon_+ = \frac{r_{\max}}{1-\gamma}\sqrt{\frac{8}{N_2}\log\left(\frac{4K(S+A)}{\delta}\right)}.$$

Note that by definitions of N_1 and N_2 we have that $\varepsilon_\square = \varepsilon_+\sqrt{\frac{K}{S+A-2K}}$. Next, recall that $\|Q_\tau^\pi - Q^\pi\|_\infty$ is upper-bounded by $\frac{r_{\max}}{T}$ from (1). Combining this with Lemma 5 and our definition of $\varepsilon_\square, \varepsilon_+$ we can see that the conditions $a)$ and $b)$ of Theorem 5 are met.

Hence, by Theorem 5 we obtain that $\|\widehat{Q}^\pi - Q^\pi\|_\infty \le \varepsilon$ if:

$$\varepsilon_+ \lesssim \frac{\varepsilon}{\|Q^\pi\|_\infty\frac{\sqrt{SA}}{\sigma_d(Q^\pi)}\log^2\left(\frac{S+A}{\delta}\right)\left(1+\|Q^\pi\|_\infty\frac{\sqrt{SA}}{\sigma_d(Q^\pi)}\log^2\left(\frac{S+A}{\delta}\right)\sqrt{\frac{K}{S+A-2K}}\right)}$$

Setting ε_+ to be equal to this value we obtain that is sufficient to have:

$$N_2 \gtrsim \frac{r_{\max}^2}{\varepsilon^2(1-\gamma)^2}\frac{\|Q^\pi\|_\infty^2 SA}{\sigma_d^2(Q^\pi)}\left(1+\frac{\|Q^\pi\|_\infty^2 SA}{\sigma_d^2(Q^\pi)}\frac{K}{S+A-2K}\log^4\left(\frac{S+A}{\delta}\right)\right)\log^5\left(\frac{d(S+A)}{\delta}\right)$$

Using definition of N_2 and rewriting inequality above in terms of T gives the following condition:

$$T \gtrsim \frac{r_{\max}^2 d}{(1-\gamma)^3\varepsilon^2}\frac{\|Q^\pi\|_\infty^2 SA}{\sigma_d^2(Q^\pi)}\left(S+A+d\frac{\|Q^\pi\|_\infty^2 SA}{\sigma_d^2(Q^\pi)}\log^5\left(\frac{S+A}{\delta}\right)\right)\log^7\left(\frac{(S+A)T}{\delta(1-\gamma)}\right)$$

Combining this with condition (12) and using the fact that $\frac{\|Q^\pi\|_\infty^2 SA}{\sigma_d^2(Q^\pi)} \le \kappa^2\alpha^2 d$ gives the final condition:

$$T = \widetilde{\Omega}_\delta\left[r_{\max}^2\kappa^2\alpha^2\frac{d(S+A)}{(1-\gamma)^3}\left(\frac{\kappa^2}{\|Q^\pi\|_\infty^2}+\frac{d}{\varepsilon^2}+\frac{d^2\alpha^2\kappa^2}{(S+A)\varepsilon^2}\right)\right]$$

Finally we verify that $\varepsilon_\square$ and ε_+ satisfy conditions from Theorem 5. Note that $\|Q^\pi\|_\infty\frac{\sqrt{SA}}{\sigma_d(Q)} \ge \|Q^\pi\|_\infty\frac{\sqrt{SA}}{\|Q^\pi\|_\mathrm{F}} \ge 1$, as well as $\log^2\left(\frac{S+A}{\delta}\right) \ge 1$ for any $\delta \in (0,1)$. Thus we obtain that $\varepsilon_+ \lesssim \varepsilon$ and thus, in order to have $\varepsilon_+ \lesssim \|Q^\pi\|_\infty$ it is sufficient to assume that $\varepsilon \lesssim \|Q^\pi\|_\infty$. Using the same reasoning we have:

$$\varepsilon_\square \lesssim \varepsilon_+\sqrt{\frac{K}{S+A}} \lesssim \varepsilon\frac{\sigma_d(Q^\pi)}{\|Q^\pi\|_\infty\sqrt{SA}\log^2\left(\frac{S+A}{\delta}\right)}\sqrt{\frac{K}{S+A}} \lesssim \frac{\sigma_d(Q^\pi)}{\sqrt{SA}}\log^{-2}\left(\frac{S+A}{\delta}\right)$$

whenever $\varepsilon \lesssim \|Q^\pi\|_\infty$ and $S+A \gtrsim K$.

C.2 Proof of Theorem 5

Note that $(L\widetilde{Q}_\tau^\pi(\mathcal{I},\mathcal{J})R)^\dagger \ne R^\dagger(\widetilde{Q}_\tau^\pi(\mathcal{I},\mathcal{J}))^\dagger L^\dagger$ in general, and thus our estimation is different from $\widetilde{Q}_\tau^\pi(s,\mathcal{J})(\widetilde{Q}_\tau^\pi(\mathcal{I},\mathcal{J}))^\dagger\widetilde{Q}_\tau^\pi(\mathcal{I},a)$ used in [35]. However, weighting estimates by inverse leverage scores as proposed in Section 4.3 still provides unbiased estimates, in the following sense:

Lemma 6. *Assume that* $|\mathcal{I}|, |\mathcal{J}| = K$ *and* $\mathrm{rank}(Q^\pi) = d$. *Then:*

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}: \quad Q^\pi(s,a) = Q^\pi(s,\mathcal{J})R(LQ^\pi(\mathcal{I},\mathcal{J})R)^\dagger LQ^\pi(\mathcal{I},a)$$

Proof of Theorem 5. The proof follows from the proof of Proposition 13 of [35], to which we refer the reader for a more detailed exposition. Based on Lemma 6 and following the proof of Proposition 13 in [35] (see (22) and (23) in [35]), we have $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$ and $(*) := |\widehat{Q}^\pi(s,a) - Q^\pi(s,a)|$:

$$(*) \leq \sqrt{2}\|(L\widetilde{Q}^\pi_\tau(\mathcal{I},\mathcal{J})R)^\dagger\|_{\text{op}}\|L(\widetilde{Q}^\pi_\tau(\mathcal{I},a)\widetilde{Q}^\pi_\tau(s,\mathcal{J}) - Q^\pi(\mathcal{I},a)Q^\pi(s,\mathcal{J}))R\|_{\text{F}} \\ + \|(L\widetilde{Q}^\pi_\tau(\mathcal{I},\mathcal{J})R)^\dagger - (LQ^\pi(\mathcal{I},\mathcal{J})R)^\dagger\|_{\text{op}}\|LQ^\pi(\mathcal{I},a)Q^\pi(s,\mathcal{J})R\|_{\text{F}} \quad (13)$$

We will repeatedly use result from Lemma 8 and condition on the event when given bounds on $\|L\|_{\text{op}}$ and $\|R\|_{\text{op}}$ hold. We begin by bounding the first term in (13). Using the assumption that $\forall(s,a) \in \mathcal{I} \times \mathcal{J}$: $|\widetilde{Q}^\pi_\tau(s,a) - Q^\pi(s,a)| \leq \varepsilon_\square$ we obtain:

$$\|L(\widetilde{Q}^\pi_\tau(\mathcal{I},\mathcal{J}) - Q^\pi(\mathcal{I},\mathcal{J}))R\|_{\text{op}} \leq \|L\|_{\text{op}}K\|\widetilde{Q}^\pi_\tau(\mathcal{I},\mathcal{J}) - Q^\pi(\mathcal{I},\mathcal{J})\|_\infty\|R\|_{\text{op}} \\ \leq c_\mathcal{I}c_\mathcal{J}\varepsilon_\square\sqrt{SA}\log^2\left(\frac{S+A}{\delta}\right).$$

Combining this inequality with our assumption on $\varepsilon_\square$ and Corollary 2 with $\eta = 1/4$ gives:

$$\|(L\widetilde{Q}^\pi_\tau(\mathcal{I},\mathcal{J})R)^\dagger\|_{\text{op}} = \frac{1}{\sigma_d(L\widetilde{Q}^\pi_\tau(\mathcal{I},\mathcal{J})R)} \\ \leq \frac{1}{\sigma_d(LQ^\pi(\mathcal{I},\mathcal{J})R) - \|L(\widetilde{Q}^\pi_\tau(\mathcal{I},\mathcal{J}) - Q^\pi(\mathcal{I},\mathcal{J}))R\|_{\text{op}}} \\ \leq \frac{8}{\sigma_d(Q^\pi)}$$

Second term in (13) can be bounded as follows:

$$\|L(\widetilde{Q}^\pi_\tau(\mathcal{I},a)\widetilde{Q}^\pi_\tau(s,\mathcal{J}) - Q^\pi(\mathcal{I},a)Q^\pi(s,\mathcal{J}))R\|_{\text{F}} \\ \leq \|L\|_{\text{op}}\|\widetilde{Q}^\pi_\tau(\mathcal{I},a)\widetilde{Q}^\pi_\tau(s,\mathcal{J}) - Q^\pi(\mathcal{I},a)Q^\pi(s,\mathcal{J})\|_{\text{F}}\|R\|_{\text{op}}$$

and then use that:

$$\|\widetilde{Q}^\pi_\tau(\mathcal{I},a)\widetilde{Q}^\pi_\tau(s,\mathcal{J}) - Q^\pi(\mathcal{I},a)Q^\pi(s,\mathcal{J})\|_{\text{F}} \leq \sqrt{|\mathcal{I}||\mathcal{J}|}(2\varepsilon_+\|Q^\pi\|_\infty + \varepsilon_+^2)$$

Combining this result with Lemma 8 we get:

$$\|L(\widetilde{Q}^\pi_\tau(\mathcal{I},a)\widetilde{Q}^\pi_\tau(s,\mathcal{J}) - Q^\pi(\mathcal{I},a)Q^\pi(s,\mathcal{J}))R\|_{\text{F}} \leq c_\mathcal{I}c_\mathcal{J}(2\|Q^\pi\|_\infty\varepsilon_+ + \varepsilon_+^2)\sqrt{SA}\log^2\left(\frac{S+A}{\delta}\right)$$

Similarly to the proof of Proposition 13 in [35], we bound the third term from 13 using inequality $\|B^\dagger - A^\dagger\|_{\text{op}} \leq \frac{1+\sqrt{5}}{2}\min\{\|A^\dagger\|^2_{\text{op}}, \|B^\dagger\|^2_{\text{op}}\}\|B - A\|_{\text{op}}$ as follows:

$$\|(L\widetilde{Q}^\pi_\tau(\mathcal{I},\mathcal{J})R)^\dagger - (LQ^\pi(\mathcal{I},\mathcal{J})R)^\dagger\|_{\text{op}} \leq 64c_\mathcal{I}c_\mathcal{J}\frac{\varepsilon_\square}{\sigma_d^2(Q^\pi)}\sqrt{SA}\log^2\left(\frac{S+A}{\delta}\right)$$

And the last term from (13) can be bounded as follows:

$$\|LQ^\pi(\mathcal{I},a)Q^\pi(s,\mathcal{J})R\|_{\text{F}} \leq \|L\|_{\text{op}}\|Q^\pi(\mathcal{I},a)Q^\pi(s,\mathcal{J})\|_{\text{F}}\|R\|_{\text{op}} \\ \leq c_\mathcal{I}c_\mathcal{J}\sqrt{SA}\|Q^\pi\|^2_\infty\log^2\left(\frac{S+A}{\delta}\right)$$

where we used that $\|Q^\pi(\mathcal{I},a)Q^\pi(s,\mathcal{J})\|_{\text{F}} \leq K\|Q^\pi\|^2_\infty$.

Combining all derived inequalities we obtain:

$$\|\widehat{Q}^\pi - Q^\pi\|_\infty \leq 8c_\mathcal{I}c_\mathcal{J}(2\|Q^\pi\|_\infty\varepsilon_+ + \varepsilon_+^2)\frac{\sqrt{SA}}{\sigma_d(Q^\pi)}\log^2\left(\frac{S+A}{\delta}\right) \\ + 64c_\mathcal{I}^2c_\mathcal{J}^2\frac{SA}{\sigma_d(Q^\pi)^2}\varepsilon_\square\|Q^\pi\|^2_\infty\log^4\left(\frac{S+A}{\delta}\right)$$

Proof of Lemma 6. First, define matrices $\mathcal{D}_U, \mathcal{D}_W \in \mathbb{R}^{K\times d}$ by $\mathcal{D}_U = LU_{\mathcal{I},:}$ and $\mathcal{D}_W = RW_{\mathcal{J},:}$, and note that $\mathcal{D}_U$ and $\mathcal{D}_W$ are not orthogonal. However, we claim and prove in the end of this proof that:

$$(\mathcal{D}_U \Sigma \mathcal{D}_W^\top)^\dagger = (\mathcal{D}_W^\top)^\dagger \Sigma^{-1} \mathcal{D}_U^\dagger \tag{14}$$

Moreover, since $\mathcal{D}_U$ and $\mathcal{D}_W$ have full column rank, we have that $\mathcal{D}_U^\dagger \mathcal{D}_U = I_{d\times d}$ and $\mathcal{D}_W^\top (\mathcal{D}_W^\top)^\dagger = I_{d\times d}$. Thus we have $\forall (s,a) \in \mathcal{S}\times\mathcal{A}$:

$$\begin{aligned} Q^\pi(s,\mathcal{J}) R (L Q^\pi(\mathcal{I},\mathcal{J}) R)^\dagger L Q^\pi(\mathcal{I},a) &= e_s^\top U \Sigma \mathcal{D}_W^\top (\mathcal{D}_U \Sigma \mathcal{D}_W^\top)^\dagger \mathcal{D}_U \Sigma W^\top e_a \\ &= e_s^\top U \Sigma (\mathcal{D}_W^\top (\mathcal{D}_W^\top)^\dagger) \Sigma^{-1} (\mathcal{D}_U^\dagger \mathcal{D}_U) \Sigma W^\top e_a \\ &= e_s^\top U \Sigma W^\top e_a = Q^\pi(s,a) \end{aligned}$$

Now we proceed with proving (14) following similar argument as in Lemma 1 in [14]. Let SVD of $\mathcal{D}_U$ and $\mathcal{D}_W$ be given by $\mathcal{D}_U = U_{\mathcal{D}_U} \Sigma_{\mathcal{D}_U} W_{\mathcal{D}_U}^\top$ and $\mathcal{D}_U = U_{\mathcal{D}_W} \Sigma_{\mathcal{D}_W} W_{\mathcal{D}_W}^\top$. First, we use that $U_{\mathcal{D}_U}$ and $U_{\mathcal{D}_W}$ are orthogonal matrices to get:

$$\begin{aligned} (\mathcal{D}_U \Sigma \mathcal{D}_W^\top)^\dagger &= (U_{\mathcal{D}_U} \Sigma_{\mathcal{D}_U} W_{\mathcal{D}_U}^\top \Sigma W_{\mathcal{D}_W} \Sigma_{\mathcal{D}_W} U_{\mathcal{D}_W}^\top)^\dagger \\ &= U_{\mathcal{D}_W} (\Sigma_{\mathcal{D}_U} W_{\mathcal{D}_U}^\top \Sigma W_{\mathcal{D}_W} \Sigma_{\mathcal{D}_W})^\dagger U_{\mathcal{D}_U}^\top \end{aligned}$$

Since $\mathcal{D}_U$ and $\mathcal{D}_W$ are matrices with full column rank, all matrices inside of the pseudoinverse are of size $d\times d$ and full rank. Thus, their product is as well full rank and replacing pseudoinverse by inverse we obtain:

$$(\mathcal{D}_U \Sigma \mathcal{D}_W^\top)^\dagger = U_{\mathcal{D}_W} \Sigma_{\mathcal{D}_W}^{-1} W_{\mathcal{D}_W}^\top \Sigma^{-1} W_{\mathcal{D}_U} \Sigma_{\mathcal{D}_U}^{-1} U_{\mathcal{D}_U}^\top = (\mathcal{D}_W^\top)^\dagger \Sigma^{-1} \mathcal{D}_U^\dagger$$

C.3 Concentration results for the proof of Theorem 5

Lemma 7. *Assume that p is a probability measure on $\mathcal{S}$ such that $p_i \geq \eta \frac{\|U_{i,:}\|_2^2}{d}$ for all $i \in [S]$ and some $\eta \in [0,1]$. Let $\mathcal{I}$ be a set obtained by sampling K entries of $\mathcal{S}$ according to p i.e. for any $i \in [S] : i \in \mathcal{I}$ with probability $\min\{1, Kp_i\}$. Define diagonal matrix L with entries $\frac{1}{\min\{1,\sqrt{Kp_i}\}}$ for $i \in \mathcal{I}$ and matrix $\mathcal{D}_U \in \mathbb{R}^{K\times d}$ given by $\mathcal{D}_U = LU_{\mathcal{I},:}$. Then, for any $\delta \in (0,1)$:*

$$\|\mathcal{D}_U^\top \mathcal{D}_U - I_{d\times d}\|_{\text{op}} \leq 2\sqrt{\frac{d}{K\eta} \log\left(\frac{2d}{\delta}\right)}$$

holds with probability at least $1-\delta$ whenever $K \geq \frac{4d}{9\eta}\log(2d/\delta)$.

Proof. First we argue that case $p_i > \frac{1}{K}$ is simple. Denote by $\mathcal{S}_+$ states for which $p_i \leq \frac{1}{K}$. Then, we have:

$$\begin{aligned} \|\mathcal{D}_U^\top \mathcal{D}_U - I_{d\times d}\|_{\text{op}} &= \|U_{\mathcal{I},:}^\top L^2 U_{\mathcal{I},:} - U^\top U\|_{\text{op}} \\ &= \left\| \sum_{i\in\mathcal{I}\cap\mathcal{S}_+} \delta_i (Z^{(i)})^\top Z^{(i)} L_{i,i}^2 - \sum_{i\in\mathcal{S}_+} (Z^{(i)})^\top Z^{(i)} \right\|_{\text{op}}, \end{aligned}$$

where $Z^{(i)}$ are obtained from U by zeroing all rows except i-th, and δ_i's are i.i.d. Bernoulli(Kp_i) for $i \in \mathcal{S}_+$. Now we can rewrite the first term:

$$\sum_{i\in\mathcal{S}_+} \delta_i^2 (Z^{(i)})^\top Z^{(i)} L_{i,i}^2 = \sum_{i\in\mathcal{S}_+} \frac{1}{Kp_i} \delta_i U_{i,:}^\top U_{i,:},$$

and take expectation over δ_i's to get $\mathbb{E}\left[\sum_{i\in\mathcal{S}_+} \delta_i^2 (Z^{(i)})^\top Z^{(i)} L_{i,i}^2\right] = \sum_{i\in\mathcal{S}_+} (Z^{(i)})^\top Z^{(i)}$.

Now, define $X^{(i)} = (\delta_i - Kp_i)\frac{1}{Kp_i} U_{i,:}^\top U_{i,:}$ for $i \in \mathcal{S}_+$. Note that:

$$\|X^{(i)}\|_{\text{op}} \leq \frac{1}{Kp_i}\|U_{i,:}\|_2^2 \leq \frac{d}{K\eta}$$

by our assumption on p. Moreover, using that $\mathrm{Var}(\delta_i) = Kp_i(1-Kp_i) \le Kp_i$ we have:

$$\mathbb{E}\left[\sum_{i\in\mathcal{S}_+} X^{(i)}(X^{(i)})^\top\right] = \sum_{i\in\mathcal{S}_+} \mathbb{E}(\delta_i - Kp_i)^2 \frac{1}{K^2p_i^2}\|U_{i,:}\|_2^2 U_{i,:}^\top U_{i,:} \le \frac{d}{K\eta}\sum_{i\in\mathcal{S}_+} U_{i,:}^\top U_{i,:}$$

implying that $\|\mathbb{E}[\sum_{i\in[S]} X^{(i)}(X^{(i)})^\top]\|_{\text{op}} \le \frac{d}{K\eta}$. Finally, noting that $X^{(i)}$ are symmetric matrices $\forall i$, we apply matrix Bernstein inequality to obtain:

$$\mathbb{P}(\|\mathcal{D}_U^\top\mathcal{D}_U - I_{d\times d}\|_{\text{op}} \ge t) = \mathbb{P}\left(\Big\|\sum_{i\in[S]} X^{(i)}\Big\|_{\text{op}} \ge t\right) \le 2d\exp\left(-\frac{K\eta}{2d}\frac{t^2}{1+\frac{t}{3}}\right)$$

Setting right hand side equal to δ finishes the proof. □

Corollary 2. *If anchor states of size at least $K \ge \frac{16d}{\eta}\log(4d/\delta)$ are chosen according to Lemma 7, we have with probability $\ge 1-\delta$:*

$$\sigma_d(LQ^\pi(\mathcal{I},\mathcal{J})R) = \sigma_d(\mathcal{D}_U\Sigma\mathcal{D}_W^\top) \ge \sigma_d(\mathcal{D}_U)\sigma_d(Q^\pi)\sigma_d(\mathcal{D}_W) \ge \frac{1}{4}\sigma_d(Q^\pi)$$

Note that we could use inequality above since $\mathcal{D}_U$ and $\mathcal{D}_W$ have full column rank.

Lemma 8. *Consider setting of Lemma 7. Then there exist universal constants $c_\mathcal{I}, c_\mathcal{J} > 0$ such that with probability at least $1-\delta$:*

$$\|L\|_{\text{op}} \le c_\mathcal{I}\sqrt{\frac{S}{K}}\log\left(\frac{S}{\delta}\right), \qquad \|R\|_{\text{op}} \le c_\mathcal{J}\sqrt{\frac{A}{K}}\log\left(\frac{A}{\delta}\right)$$

Proof. We note that if $L_{i,i} = 1$ (i.e. $Kp_i \ge 1$), then obviously the inequality above holds for $S \ge K$, and thus we consider only cases where $L_{i,i} = \frac{1}{\sqrt{Kp_i}}$. Now, note that $\|L\|_{\text{op}} = \|L^{\text{ext}}\|_{\text{op}}$, where $L^{\text{ext}} = \sum_{i=1}^S \delta_i \frac{1}{\sqrt{Kp_i}} e_i e_i^\top$, and where δ_i are i.i.d. Bernoulli(Kp_i). Next, we have: $\mathbb{E}[\delta_i \frac{1}{\sqrt{Kp_i}} e_i e_i^\top] = \sqrt{Kp_i}e_i e_i^\top$, and thus:

$$\|\mathbb{E}L^{\text{ext}}\|_{\text{op}} = \sqrt{K\max_i p_i} \le \sqrt{K}$$

Define $Y^{(i)} = (\delta_i - Kp_i)\frac{1}{\sqrt{Kp_i}} e_i e_i^\top$. We have $\mathbb{E}[Y^{(i)}(Y^{(i)})^\top] = (1-Kp_i)e_i e_i^\top$, and hence the variance term in matrix Bernstein is upper bounded by 1. Lastly by our assumption on p we have for all $i \in [S]$:

$$\|Y^{(i)}\|_{\text{op}} \le \frac{1}{\sqrt{Kp_i}} \le c\sqrt{\frac{S}{K}}$$

By matrix Bernstein we obtain:

$$\mathbb{P}(\|L^{\text{ext}} - \mathbb{E}L^{\text{ext}}\|_{\text{op}} \ge t) = \mathbb{P}\left(\Big\|\sum_{i\in[S]} Y^{(i)}\Big\|_{\text{op}} \ge t\right) \le 2S\exp\left(-\frac{\frac{t^2}{2}}{1+t\frac{c}{3}\sqrt{\frac{S}{K}}}\right)$$

Equating last term with δ and using that $S, A \gg d$ we obtain statement of the lemma. □

D Sample Complexity Analysis of LoRa-PI

In this appendix, we present the proof of Theorem 3. It is a direct consequence of the performance guarantee of LME (see Theorem 2) and an error bound on approximate policy iteration, which we provide in this appendix (see Lemma 9).

D.1 Proof of Theorem 3

Proof of Theorem 3. To start with, we first observe that, according to Lemma 9, LoRa-PI outputs $\hat{\pi}$ with $\|V^\star - V^{\hat{\pi}}\|_\infty \leq \varepsilon$, if it holds that

$$(i) \qquad \gamma^{t-1}\|V^\star - V^{\pi^{(1)}}\|_\infty \leq \frac{2r_{\max}\gamma^{(N_{\text{epochs}})}}{1-\gamma} \leq \frac{\varepsilon}{2}$$

$$(ii) \qquad \|\widehat{Q}^{(t)} - Q^{(t)}\|_\infty \leq \frac{(1-\gamma)^2\varepsilon}{4}, \quad \forall t \in [N_{\text{epochs}}]$$

where we introduce the notation $Q^{(t)} := Q^{\pi^{(t)}}$ as a shorthand. Now, we note that condition *(i)* is satisfied if

$$N_{\text{epochs}} = \left\lceil \frac{1}{1-\gamma} \log\left(\frac{4r_{\max}}{(1-\gamma)\varepsilon}\right)\right\rceil$$

which is already as chosen in LoRa-PI. Now, in order for (ii) to hold we use Theorem 2. We define the events:

$$\mathcal{E}_t = \left\{\|\widehat{Q}^{(t)} - Q^{(t)}\|_\infty \leq \frac{(1-\gamma)^2\varepsilon}{4}\right\}$$

We show that $\cap_{t\in[N_{\text{epochs}}]}\mathcal{E}_t$ holds with high probability. To that end, it is sufficient to analyse for each $t \in [N_{\text{epochs}}]$, the event $\mathcal{E}_t^c$ conditionally on the event that $(\cap_{k\in[t-1]}\mathcal{E}_k)$ holds. Indeed, by using the elementary inequality $\mathbb{P}(\mathcal{E}^c \cup \mathcal{B}^c) \leq \mathbb{P}(\mathcal{E}^c|\mathcal{B}) + \mathbb{P}(\mathcal{B}^c)$ in a recursive manner, we can write

$$\mathbb{P}((\cap_{t\in[N_{\text{epochs}}]}\mathcal{E}_t)^c) = \mathbb{P}(\cup_{t\in[N_{\text{epochs}}]}\mathcal{E}_t^c) \leq \sum_{t\in[N_{\text{epochs}}]} \mathbb{P}(\mathcal{E}_t^c|\cap_{k\in[t-1]}\mathcal{E}_k) \tag{15}$$

We will show that for all $t \in [N_{\text{epochs}}]$, $\mathbb{P}(\mathcal{E}_t^c|\cap_{k\in[t-1]}\mathcal{E}_k) \leq \delta/N_{\text{epochs}}$, which would entail that $\mathbb{P}((\cap_{t\in[N_{\text{epochs}}]}\mathcal{E}_t)^c) \leq \delta$ and ensure that $\|V^\star - V^{\hat{\pi}}\|_\infty \leq 1-\varepsilon$ holds with probability at least $1-\delta$.

Let $t \in [N_{\text{epochs}}]$. Note that by using Theorem 2, we can immediately show that $\mathbb{P}(\mathcal{E}_t^c|\cap_{k\in[t-1]}\mathcal{E}_k) \leq \delta/N_{\text{epochs}}$ provided that

$$\frac{T}{N_{\text{epochs}}} = \widetilde{\Omega}\left(\frac{r_{\max}^2\kappa^4\alpha^2d^2\left((S+A)+\alpha^2 d\right)}{(1-\gamma)^7\varepsilon^2}\ \log^{10}\left(\frac{N_{\text{epochs}}}{\delta}\right)\right)$$

which entails, by definition of N_{epochs} as chosen in LoRa-PI, an equivalent sample complexity to

$$\begin{aligned} T &= \widetilde{\Omega}\left(\frac{r_{\max}^2\kappa^4\alpha^2d^2\left((S+A)+\alpha^2 d\right)}{(1-\gamma)^8\varepsilon^2}\ \log^{10}\left(\frac{1}{(1-\gamma)\delta}\log\left(\frac{r_{\max}}{(1-\gamma)\varepsilon}\right)\right)\log\left(\frac{r_{\max}}{(1-\gamma)\varepsilon}\right)\right) \\ &= \widetilde{\Omega}\left(\frac{r_{\max}^2\kappa^4\alpha^2d^2\left((S+A)+\alpha^2 d\right)\log^{10}(e/\delta)\log(e/\varepsilon)}{(1-\gamma)^8\varepsilon^2}\right) \end{aligned}$$

where we emphasize that $\widetilde{\Omega}(\cdot)$ may hide poly-log dependencies on S, A, $(1-\gamma)^{-1}$, d, κ, α, $r_{\max}$, $\log(e/\varepsilon)$, $\log(e/\delta)$. This the desired sample complexity in Theorem 3.

Note that Theorem 2 also requires that

$$\frac{(1-\gamma)^2\varepsilon}{4} \leq \|Q^{(t)}\|_\infty \tag{16}$$

We show that this is satisfied by the condition $\varepsilon \lesssim \underline{\varepsilon}$. First, we show that under this condition, we have $\|Q^{(1)}\| \leq 2\|Q^{(t)}\|_\infty$. Using Lemma 10, and conditionally on the event $\cap_{k\in[t]}\mathcal{E}_k$ holding, we have that for all $k < t$,

$$\mathcal{T}^\star(V^{(k)}) \leq V^{(k+1)} + \frac{2\epsilon}{1-\gamma}\mathbf{1} \leq \mathcal{T}^\star\left(V^{(k+1)} + \frac{2\varepsilon^{(k)}}{1-\gamma}\mathbf{1}\right) \leq \mathcal{T}^\star(V^{(k+1)}) + \frac{2\gamma\epsilon}{1-\gamma}$$

where $\epsilon = \frac{(1-\gamma)^2\varepsilon}{4}$, implying, in particular, that

$$\|Q^{(k)}\|_\infty \leq \|Q^{(k+1)}\|_\infty + \frac{2\gamma(1-\gamma)\varepsilon}{2}.$$

Summing the above inequalities from 1 to $t-1$, together with the fact that $t - 1 \leq N_{\text{epochs}}$, gives

$$\|Q^{(1)}\|_\infty \leq \|Q^{(t)}\|_\infty + \frac{\gamma(1-\gamma)(t-1)\varepsilon}{2} \leq \|Q^{(t)}\|_\infty + \frac{\gamma\varepsilon}{2}\log\left(\frac{4r_{\max}}{(1-\gamma)^2\varepsilon}\right).$$

In view of this inequality, we note that $2\|Q^{(t)}\|_\infty \geq \|Q^{(1)}\|_\infty$, if

$$\gamma\varepsilon\log\left(\frac{4r_{\max}}{(1-\gamma)^2\varepsilon}\right) \leq \|Q^{(1)}\|_\infty$$

We can verify that the above condition is implied by:

$$\frac{1}{\varepsilon} \geq \frac{4\gamma}{\|Q^{(1)}\|_\infty}\log\left(\frac{16\gamma r_{\max}}{(1-\gamma)^2\|Q^{(1)}\|_\infty}\right) \iff \varepsilon \leq \frac{\|Q^{(1)}\|_\infty}{2\gamma\log\left(\frac{16\gamma r_{\max}}{(1-\gamma)^2\|Q^{(1)}\|_\infty}\right)} \tag{17}$$

where we used the elementary fact $x \geq 2a\log(2a) + 2b \implies x \geq a\log(x) + b$ for all $a, b > 0$.

Thus, from (17) we conclude that the condition on ε, (16), is satisfied if the following condition holds:

$$\varepsilon \leq \min\left(1, \frac{1}{2\gamma\log\left(\frac{16\gamma r_{\max}}{(1-\gamma)^2\|Q^{(1)}\|_\infty}\right)}\right)\|Q^{(1)}\|_\infty.$$

This is the desired condition on ε in Theorem 3. With this we have concluded the proof. □

D.2 Error Bound for Approximate Policy Iteration

The following result, a standard variant of Proposition 6.2 in [4], shows that the described approximate policy iteration is guaranteed to converge within an ϵ-accuracy.

Lemma 9. *Let $(\pi^{(t)})_{t\geq 1}$ be a sequence of deterministic policies selected recursively as described in* `LoRa-PI`*, and denote $V^{(t)} = V^{\pi^{(t)}}$ for all $t \geq 1$. Let $\epsilon > 0$ and suppose that for all $t \geq 1$, it holds that*

$$\|\widehat{Q}^{(t)} - Q^{(t)}\|_\infty \leq \epsilon.$$

Then, for all $t \geq 1$, we have

$$\|V^\star - V^{(t+1)}\| \leq \gamma^t\|V^\star - V^{(1)}\|_\infty + \frac{2\epsilon}{(1-\gamma)^2}.$$

The proof of Lemma 9 follows standard arguments, but we provide it for completeness.

Lemma 10. *Let π be a deterministic policy, and assume that $\|\widehat{Q}^\pi - Q^\pi\|_\infty \leq \epsilon$. Assume that policy π' is selected greedily with respect to $\widehat{Q}^\pi$, i.e., for all $s \in \mathcal{S}$, $\pi'(s) = \arg\max_{a\in\mathcal{A}}\widehat{Q}^\pi(s,a)$, then*

$$V^\pi \leq \mathcal{T}^\star(V^\pi) \leq V^{\pi'} + \frac{2\epsilon}{1-\gamma}\mathbf{1}.$$

Proof of Lemma 10. Before we proceed with the proof, let us define the composition of a deterministic policy π'' and a Q^π function, $\pi'' \circ Q^\pi(s) := Q^\pi(s, \pi''(s))$. We know that

$$V^\pi = \pi \circ Q^\pi \leq \max_{\pi''} \pi'' \circ Q^\pi = \mathcal{T}^\star(V^\pi)$$

where $\leq$ is applied component-wise. Next, we have

$$\begin{aligned} V^\pi \leq \mathcal{T}^\star(V^\pi) = \max_{\pi''} \pi'' \circ Q^\pi &\leq \max_{\pi''} \pi'' \circ \widehat{Q}^\pi + \max_{\pi''} \pi'' \circ (Q^\pi - \widehat{Q}^\pi) \\ &\leq \pi' \circ \widehat{Q}^\pi + \epsilon \mathbf{1} \\ &\leq \pi' \circ Q^\pi + \pi' \circ (\widehat{Q}^\pi - Q^\pi) + \epsilon \mathbf{1} \\ &\leq \pi' \circ Q^\pi + 2\epsilon \mathbf{1} \\ &\leq \mathcal{T}_{\pi'}(V^\pi) + 2\varepsilon \mathbf{1} \end{aligned}$$

where $\mathcal{T}_\pi$ is the Bellman policy evaluation operator. By monotonicity of the operator $\mathcal{T}_\pi$, we can re-iterate

$$\mathcal{T}_{\pi'}(V^\pi) \leq \mathcal{T}_{\pi'}(\mathcal{T}_{\pi'}(V^\pi) + 2\epsilon \mathbf{1}) \leq \mathcal{T}_{\pi'}^2(V^\pi) + 2\gamma\epsilon \mathbf{1}.$$

Thus, we finally obtain

$$V^\pi \leq \mathcal{T}^\star(V^\pi) \leq \mathcal{T}_{\pi'}^{k+1}(V^\pi) + 2\epsilon \left(\sum_{t=0}^{k} \gamma^t \right) \mathbf{1}.$$

Taking $k \to \infty$, we get

$$V^\pi \leq \mathcal{T}^\star(V^\pi) \leq \mathcal{T}_{\pi'}(V^\pi) + 2\epsilon \mathbf{1} \leq V^{\pi'} + \frac{2\epsilon}{1-\gamma} \mathbf{1}.$$

□

Proof of Lemma 9. We start by noting that, thanks to Lemma 10, we have: for all $t \geq 1$,

$$V^{(t+1)} + \frac{2\epsilon}{1-\gamma} \mathbf{1} \geq \mathcal{T}^\star(V^{(t)}),$$

where $\geq$ is applied component-wise. Thus, applying this inequality recursively we obtain

$$\begin{aligned} V^{(t+1)} + \frac{2\epsilon}{1-\gamma} \mathbf{1} &\geq \mathcal{T}^\star \left(V^{(t)} + \frac{2\epsilon}{1-\gamma} \mathbf{1} \right) - \frac{2\epsilon\gamma}{1-\gamma} \mathbf{1} \\ &\geq (\mathcal{T}^\star)^2 \left(V^{(t-1)} \right) - \frac{2\epsilon\gamma}{1-\gamma} \mathbf{1} \\ &\geq (\mathcal{T}^\star)^t (V^{(1)}) - \frac{2\epsilon}{1-\gamma} \left(\sum_{k=1}^{t-1} \gamma^k \right) \mathbf{1} \\ &\geq (\mathcal{T}^\star)^t (V^{(1)}) - \frac{2\epsilon(1-\gamma^t)}{(1-\gamma)^2} \mathbf{1} + \frac{2\epsilon}{1-\gamma} \mathbf{1}, \end{aligned}$$

which gives at the end

$$V^{(t+1)} \geq (\mathcal{T}^\star)^{(t)} (V^{(1)}) - \frac{2\epsilon(1-\gamma^t)}{(1-\gamma)^2} \mathbf{1}$$

Thus, we have

$$V^\star - V^{(t+1)} \leq V^\star - (\mathcal{T}^\star)^t (V^{(1)}) + \frac{2\epsilon(1-\gamma^t)}{(1-\gamma)^2} \mathbf{1} \leq (\mathcal{T}^\star)^t (V^\star) - (\mathcal{T}^\star)^t (V^{(1)}) + \frac{2\epsilon(1-\gamma^t)}{(1-\gamma)^2} \mathbf{1}$$

Thus, using the contraction property of $\mathcal{T}^\star$, and that $1 - \gamma^t \leq 1$, we have

$$\|V^\star - V^{(t+1)}\|_\infty \leq \|(\mathcal{T}^\star)^t (V^\star) - (\mathcal{T}^\star)^t (V^{(1)})\|_\infty + \frac{2\epsilon}{(1-\gamma)^2} \leq \gamma^t \|V^\star - V^{(1)}\|_\infty + \frac{2\epsilon}{(1-\gamma)^2}.$$

□

E Extension of Guarantees to Approximately-Low Rank MDPs

We consider the setting where the matrix Q^π is approximately low rank. Specifically, we define a constant ζ_d such that $\zeta_d = \|Q^\pi(s,a) - Q_d^\pi(s,a)\|_\infty$, where Q_d^π is the best d-rank approximation of Q^π in the operator norm. Note that $\zeta_d \leq \sigma_{d+1}(Q^\pi) \leq \sqrt{SA}\zeta_d$. In contrast to Theorem 4, where the additional perturbation term Δ arises from a controllable quantity (through roll-out length τ), here we assume that ζ_d is fixed in advance and unknown. For simplicity, we omit terms stemming from the Δ perturbation, but the results still hold in that setting. Here, we show that if:

$$\zeta_d = \widetilde{O}\left(\sigma_d(Q^\pi)\min\left\{\frac{\sqrt{d}}{S+A}, \frac{1}{\kappa\sqrt{SA}}\right\}\right) \tag{A_+}$$

we can obtain similar guarantees for $\|V^\star - V^{\hat{\pi}}\|_\infty$ as in Theorem 3 even in the approximate low rank setting, with an additive error scaling with $\widetilde{O}(\frac{1}{1-\gamma}\zeta_d d\kappa^2\alpha^2)$. Next, we show that our three main theorems still hold in this setting.

Theorem 1: Leverage scores estimation. We can repeat the arguments from the proof of Theorem 4 to obtain, with high probability, $\forall s \in [S]$:

$$\|U_{s,:} - \widehat{U}_{s,:}O_{\widehat{U}}\|_2 = \widetilde{O}\left(\bar{\alpha}\left(\sqrt{\frac{d}{T_\tau}} + \kappa\|U_{s,:}\|_2\sqrt{\frac{S+A}{T_\tau}}\right) + \frac{\zeta_d\sqrt{S+A}}{\sigma_d(Q^\pi)} + \kappa\|U_{s,:}\|_2\frac{\sigma_{d+1}(Q^\pi)}{\sigma_d(Q^\pi)}\right)$$

if $T_\tau = \widetilde{\Omega}\left(\bar{\alpha}^2(S+A)\right)$, and $\sigma_{d+1}(Q^\pi) \leq \sigma_d(Q^\pi)/64$. New terms are highlighted in blue in the inequality above. A similar inequality holds for the rows of the matrix of right singular vectors W.

Under Assumption A_+ and using that $\sigma_{d+1}(Q^\pi) \leq \sqrt{SA}\zeta_d$, we have:

$$\frac{\zeta_d\sqrt{S+A}}{\sigma_d(Q^\pi)} = \widetilde{O}\left(\frac{\sqrt{d}}{\sqrt{S+A}}\right), \qquad \text{and} \qquad \kappa\|U_{s,:}\|_2\frac{\sigma_{d+1}(Q^\pi)}{\sigma_d(Q^\pi)} = \widetilde{O}\left(\|U_{s,:}\|_2\right)$$

indicating that the contributions of the two newly added terms are negligible for leverage score estimation and that Theorem 1 still holds in this setting.

Theorem 2: Complete matrix estimation. Theorem 5 holds with the same arguments. Instead of Lemma 5, we have that with high probability: $\forall(s,a) \in (\mathcal{I}\times\mathcal{A}) \cup (\mathcal{S}\times\mathcal{J})$:

$$|\widetilde{Q}_\tau^\pi(s,a) - Q^\pi(s,a)| \leq \frac{r_{\max}}{1-\gamma}\sqrt{\frac{2}{N}\log\left(\frac{4K(S+A)}{\delta}\right)} + \zeta_d$$

Note that our conditions on ζ_d and $\sigma_{d+1}(Q^\pi)$ ensure that the conditions on $\varepsilon_\square$ and ε_+ in Theorem 5 ($\varepsilon_\square \lesssim \frac{\sigma_d(Q^\pi)}{\sqrt{SA}}\log^{-2}\left(\frac{S+A}{\delta}\right)$, $\varepsilon_+ \lesssim \|Q^\pi\|_\infty$) still hold, as:

$$\zeta_d = \widetilde{O}\left(\frac{\sqrt{d}\sigma_d(Q^\pi)}{S+A}\right) = \widetilde{O}\left(\frac{\|Q^\pi\|_{\mathrm{F}}}{S+A}\right) = \widetilde{O}\left(\frac{\sqrt{SA}\|Q^\pi\|_\infty}{S+A}\right) = \widetilde{O}\left(\|Q^\pi\|_\infty\right)$$

Then, the upper bound on $\|\widehat{Q}^\pi - Q^\pi\|_\infty$ from Theorem 5 will include an additive term: $\zeta_d\frac{SA\|Q^\pi\|_\infty^2}{\sigma_d^2(Q^\pi)}\log^4\left(\frac{S+A}{\delta}\right) = \widetilde{O}\left(\zeta_d d\kappa^2\alpha^2\right)$. Finally, under approximate low-rank structure, Theorem 2 guarantees that with high probability, if $\varepsilon \lesssim \|Q^\pi\|_\infty$ and $T = \widetilde{\Omega}_\delta\left(\frac{(S+A)+\alpha^2 d}{(1-\gamma)^3\varepsilon^2}(r_{\max}^2\kappa^4\alpha^2 d^2)\right)$, we have $\|\widehat{Q}^\pi - Q^\pi\|_\infty \leq \varepsilon + \widetilde{O}\left(\zeta_d d\kappa^2\alpha^2\right)$.

This aligns with Theorem 14 in [34], where the approximation error scales by terms corresponding to $\frac{SA\|Q^\pi\|_\infty^2}{\sigma_d^2(Q^\pi)}$ in our setting, as both methods use CUR-like matrix recovery.

Theorem 3: Guarantee for `LoRa-PI`. Based on the approximate policy iteration theorem, which claims:

$$(1-\gamma)\|V^\star - V^{\hat{\pi}}\|_\infty \leq 2r_{\max}\gamma^{N_{\text{epochs}}} + 2\max_{t\in[N_{\text{epochs}}]}\|\widehat{Q}^{(t)} - Q^{\pi^{(t)}}\|_\infty.$$

we observe that the error from approximate low rank propagates through the second term, yielding an additive error of magnitude $\frac{1}{1-\gamma}\zeta_d d\kappa^2\alpha^2$ to the error of Theorem 3.

F Miscellaneous Results

In this section, we provide some of the observations and results about the truncated value matrix. More specifically, we present the proof to Lemma 1, and a discussion on the variance proxy of the truncated discounted sum of rewards.

F.1 Truncated Value Matrix

Proof of Lemma 1. We know that Q^π satisfies the following identity: for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t^\pi \middle| (s_0^\pi, a_0^\pi) = (s,a)\right] \\ &= Q_\tau^\pi(s,a) + \gamma^\tau \mathbb{E}\left[\sum_{t=\tau+1}^{\infty} \gamma^{t-\tau} r_t^\pi \middle| (s_0^\pi, a_0^\pi) = (s,a)\right]. \end{aligned}$$

Furthermore, note that

$$\left|\mathbb{E}\left[\sum_{t=\tau+1}^{\infty} \gamma^{t-\tau} r_t^\pi \middle| (s_0^\pi, a_0^\pi) = (s,a)\right]\right| \leq r_{\max} \sum_{t=0}^{\infty} \gamma^t = \frac{r_{\max}}{1-\gamma}.$$

and thus

$$\|Q_\tau^\pi - Q^\pi\|_\infty \leq \frac{\gamma^\tau r_{\max}}{1-\gamma} \leq \frac{r_{\max}}{1-\gamma} \exp(-\tau(1-\gamma)).$$

where we used that $\gamma \leq \exp(\gamma - 1)$ for $\gamma \in (0,1)$. Setting the right hand side of the last inequality equal to ϵ we obtain statement of the lemma.

□

F.2 Equivalent Noise Model

Recall definition of $\widetilde{Q}_\tau^\pi$ from (2):

$$\widetilde{Q}_\tau^\pi(s,a) = \frac{SA}{N} \sum_{k=1}^{N} \left(\sum_{t=0}^{\tau} \gamma^t r_{k,t}^\pi\right) \mathbb{1}_{\{(s_{k,0}^\pi, a_{k,0}^\pi) = (s,a)\}}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Consider one of N sampled trajectories with index k starting from $(s_{k,0}^\pi, a_{k,0}^\pi) = (s,a)$, and note that

$$\left|\sum_{t=0}^{\tau} \gamma^t r_{k,t}^\pi\right| \leq \frac{r_{\max}}{1-\gamma}.$$

Moreover, since Q_τ^π is given by:

$$Q_\tau^\pi(s,a) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t \mathbb{1}_{\{t \leq \tau\}} \middle| s_0^\pi = s, a_0^\pi = a\right],$$

we have

$$\mathbb{E}\left[\sum_{t=0}^{\tau} \gamma^t r_{k,t}^\pi \mathbb{1}_{\{(s_{k,0}^\pi, a_{k,0}^\pi) = (s,a)\}} \middle| (s_{k,0}^\pi, a_{k,0}^\pi) = (s,a)\right] = Q_\tau^\pi(s,a) \mathbb{1}_{\{(s_{k,0}^\pi, a_{k,0}^\pi) = (s,a)\}}$$

In other words, each term inside of the outer loop in definition of $\widetilde{Q}_\tau^\pi$ is uniformly bounded and equal to $Q_\tau^\pi(s,a)\mathbb{1}_{\{(s_{k,0}^\pi, a_{k,0}^\pi) = (s,a)\}}$ in expectation. Thus we can view estimate $\widetilde{Q}_\tau^\pi(s,a)$ equivalently as:

$$\widetilde{Q}_\tau^\pi(s,a) = \frac{SA}{N} \sum_{k=1}^{N} (Q_\tau^\pi(s,a) + \xi_{s,a,k}) \mathbb{1}_{\{(s_{k,0}^\pi, a_{k,0}^\pi) = (s,a)\}}$$

where $\xi_{s,a,k}$ are i.i.d. across k, and $|Q^\pi_\tau(s,a) + \xi_{s,a,k}| \leq \frac{r_{\max}}{1-\gamma}$, implying that $\xi_{s,a,k}$ are $\frac{2r_{\max}}{1-\gamma}$-subgaussian random variables.

Next note that the number of times $N(s,a) = \sum_{k=1}^{N} \mathbb{1}\{(s^\pi_{k,0}, a^\pi_{k,0}) = (s,a)\}$ that we sample entries are random variables with multinomial distribution, since $\mathbb{P}((s^\pi_{k,0}, a^\pi_{k,0}) = (s,a)) = \frac{1}{SA}$ and $\sum_{(s,a)} N(s,a) = N$. This weak dependence between the entries can be dealt with using the Poisson approximation argument (see Section C.2 in [37]). Essentially, this enables us to rewrite matrix $\widetilde{Q}^\pi_\tau$ as a matrix with i.i.d. entries. Namely, we have for all (s,a):

$$\widetilde{Q}^\pi_\tau(s,a) = \frac{SA}{N} \sum_{k=1}^{Y(s,a)} (Q^\pi_\tau(s,a) + \xi_{s,a,k})$$

where $Y(s,a)$ are i.i.d. Poisson random variables with parameter $\mathbb{E}[Y(s,a)] = \frac{N}{SA}$. The fact that the two noise models are equivalent is depicted in Lemma 20 in [37] claiming that probability of an event under the multinomial model can be upper bounded by $\sqrt{T}$ times probability of the same event under the Poisson model. Practically, this adds a multiplicative factor of T in our probabilistic claims. For more thorough exposition of this issue check Section C.2 in [37].