Offline Primal-Dual Reinforcement Learning for Linear MDPs

Germano Gabbianelli
Universitat Pompeu Fabra
Barcelona, Spain

Gergely Neu
Universitat Pompeu Fabra
Barcelona, Spain

Nneka Okolo
Universitat Pompeu Fabra
Barcelona, Spain

Matteo Papini
Politecnico di Milano,
Milan, Italy

Abstract

Offline Reinforcement Learning (RL) aims to learn a near-optimal policy from a fixed dataset of transitions collected by another policy. This problem has attracted a lot of attention recently, but most existing methods with strong theoretical guarantees are restricted to finite-horizon or tabular settings. In contrast, few algorithms for infinite-horizon settings with function approximation and minimal assumptions on the dataset are both sample and computationally efficient. Another gap in the current literature is the lack of theoretical analysis for the average-reward setting, which is more challenging than the discounted setting. In this paper, we address both of these issues by proposing a primal-dual optimization method based on the linear programming formulation of RL. Our key contribution is a new reparametrization that allows us to derive low-variance gradient estimators that can be used in a stochastic optimization scheme using only samples from the behavior policy. Our method finds an ε-optimal policy with $O(\varepsilon^{-4})$ samples, while being computationally efficient for infinite-horizon discounted and average-reward MDPs with realizable linear function approximation and partial coverage. Moreover, to the best of our knowledge, this is the first theoretical result for average-reward offline RL.

1 INTRODUCTION

We study the setting of Offline Reinforcement Learning (RL), where the goal is to learn an ε-optimal policy without being able to interact with the environment, but only using a fixed dataset of transitions collected by a *behavior policy*. Learning from offline data proves to be useful especially when interacting with the environment can be costly or dangerous (Levine et al., 2020).

In this setting, the quality of the best policy learnable by any algorithm is constrained by the quality of the data, implying that finding an optimal policy without further assumptions on the data is not feasible. Therefore, many methods (Munos and Szepesvári, 2008; Uehara et al., 2020) make a *uniform coverage* assumption, requiring that the behavior policy explores sufficiently well the whole state-action space. However, recent work (Liu et al., 2020; Rashidinejad et al., 2022) demonstrated that *partial coverage* of the state-action space is sufficient. In particular, this means that the behavior policy needs only to sufficiently explore the state-action pairs visited by the optimal policy.

Moreover, like its online counterpart, modern offline RL faces the problem of learning efficiently in environments with very large state spaces, where function approximation is necessary to compactly represent policies and value functions. Although function approximation, especially with neural networks, is widely used in practice, its theoretical understanding in the context of decision-making is still rather limited, even when considering *linear* function approximation.

In fact, most existing sample complexity results for offline RL algorithms are limited either to the tabular and finite horizon setting, by the uniform coverage assumption or by assuming access to a (convex) optimization oracle — see the top section of Table 1 for a summary. Notable exceptions in terms of computational efficiency are the works of Xie et al. (2021) and Cheng et al. (2022), who provide a computationally efficient version of their method for infinite-horizon discounted MDPs under realizable linear function approximation and partial coverage assumptions. Despite being some of the first concrete implementations, the practical versions of those algorithms differ significantly from their information-theoretic counterparts, and thus the sample-complexity guarantees proven in

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

the corresponding papers do not immediately carry over to them.

More similar to our work are those of Zhan et al. (2022), and Rashidinejad et al. (2023) who also consider a linear programming approach to offline learning in infinite-horizon discounted MDPs. Yet, like many works which consider the broader general function approximation setting, their method may remain oracle-efficient even in the simpler linear MDP setting – see the caption of Table 1. Moreover, all methods referenced so far only work in the finite-horizon or infinite-horizon *discounted* setting, which is inappropriate for modeling practical problems where it is hard to pre-specify a fixed decision-making horizon. This issue is readily addressed by the average-reward framework, which however is known to be much more difficult to handle using techniques familiar from the discounted-reward setting. For example, methods based on approximate dynamic programming like Zhu et al. (2023) make crucial use of the contractive property of the discounted Bellman operators, which does not generally hold in the average-reward setting (especially not under the general assumptions we make in our work). Therefore, this work is motivated by the following research question:

Can we design a linear-time algorithm with polynomial sample complexity for the discounted and average-reward infinite-horizon settings, in large state spaces under a partial-coverage assumption?

We answer this question positively by designing a method based on the linear-programming (LP) formulation of sequential decision making (Manne, 1960a). Albeit less known than the dynamic-programming formulation (Bellman, 1956) that is ubiquitous in RL, it allows us to tackle this problem with the powerful tools of convex optimization. We turn in particular to a relaxed version of the LP formulation (Mehta and Meyn, 2009; Bas-Serrano et al., 2021) that considers action-value functions that are linear in known state-action features. This allows to reduce the dimensionality of the problem from the cardinality of the state space to the number of features. This relaxation still allows to recover optimal policies in *linear MDPs* (Yang and Wang, 2019; Jin et al., 2020), a structural assumption that is widely employed in the theoretical study of RL with linear function approximation.

Our algorithm for learning near-optimal policies from offline data is based on primal-dual optimization of the Lagrangian of the relaxed LP. The use of saddle-point optimization in MDPs was first proposed by Wang and Chen (2016) for *planning* in small state spaces, and was extended to linear function approximation by Chen et al. (2018); Bas-Serrano and Neu (2020), and Neu and Okolo (2023). We largely take inspiration from this latter work, which was the first to apply saddle-point optimization to the *relaxed* LP. However, primal-dual planning algorithms assume oracle access to a transition model, whose samples are used to estimate gradients. In our offline setting, we only assume access to i.i.d. samples generated by a possibly unknown behavior policy. To adapt the primal-dual optimization strategy to this setting we employ a change of variable, inspired by Nachum and Dai (2020), which allows easy computation of unbiased gradient estimates.

Notation. We denote vectors with bold letters, such as $\boldsymbol{x} \doteq [x_1, \dots, x_d]^\top \in \mathbb{R}^d$, and use $\boldsymbol{e}_i$ to denote the i-th standard basis vector. We interchangeably denote functions $f : \mathcal{X} \to \mathbb{R}$ over a finite set $\mathcal{X}$, as vectors $\boldsymbol{f} \in \mathbb{R}^{|\mathcal{X}|}$ with components $f(x)$, and use $\geq$ to denote element-wise comparison. We denote the set of probability distributions over a measurable set $\mathcal{S}$ as $\Delta_{\mathcal{S}}$, and the probability simplex in $\mathbb{R}^d$ as Δ_d. For a policy $\pi : \mathcal{X} \to \Delta_{\mathcal{A}}$ and function $\nu : \mathcal{X} \to \mathbb{R}$ (or corresponding vector $\boldsymbol{\nu} \in \mathbb{R}^{|\mathcal{X}|}$), we use the notation $(\pi \circ \boldsymbol{\nu})(x, a) = \pi(a|x)\nu(x)$. We use $\sigma : \mathbb{R}^d \to \Delta_d$ to denote the softmax function defined as $\sigma_i(\boldsymbol{x}) \doteq e^{x_i} / \sum_{j=1}^d e^{x_j}$. We use upper-case letters for random variables, such as S, and denote the uniform distribution over a finite set of n elements as $\mathcal{U}(n)$. In the context of iterative algorithms, we use $\mathcal{F}_{t-1}$ to denote the sigma-algebra generated by all events up to the end of iteration $t-1$, and use the shorthand notation $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_{t-1}]$ to denote expectation conditional on the history. For nested-loop algorithms, we write $\mathcal{F}_{t,i-1}$ for the sigma-algebra generated by all events up to the end of iteration $i-1$ of round t, and $\mathbb{E}_{t,i}[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_{t,i-1}]$ for the corresponding conditional expectation.

2 PRELIMINARIES

We study discounted Markov decision processes (MDP, Puterman, 1994) denoted as $(\mathcal{X}, \mathcal{A}, p, r, \gamma)$, with discount factor $\gamma \in [0, 1]$ and finite, but potentially very large, state space $\mathcal{X}$ and action space $\mathcal{A}$. For every state-action pair (x, a), we denote as $p(\cdot \mid x, a) \in \Delta_{\mathcal{X}}$ the next-state distribution, and as $r(x, a) \in [0, 1]$ the reward, which is assumed to be deterministic and bounded for simplicity. The transition function p is also denoted as the matrix $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{X}\times\mathcal{A}|\times|\mathcal{X}|}$ and the reward as the vector $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{X}\times\mathcal{A}|}$. The objective is to find an *optimal policy* $\pi^* : \mathcal{X} \to \Delta_{\mathcal{A}}$. That is, a stationary policy that maximizes the normalized expected return $\rho(\pi^*) \doteq (1-\gamma)\mathbb{E}_{\pi^*}[\sum_{t=0}^\infty \gamma^t r(X_t, A_t)]$, where the initial state X_0 is sampled from the initial state distribution ν_0, the other states according to $X_{t+1} \sim p(\cdot|X_t, A_t)$ and where the notation $\mathbb{E}_\pi[\cdot]$ is used to denote that the actions are sampled from policy

Algorithm	Partial Coverage	Sample Complexity	Computational Complexity	Function Approximation	Infinite Horizon	
					Discounted	Avg. Reward
PEVI (Jin et al., 2021)	✓	$O(\varepsilon^{-2})$	$O(n)$	general	✗	✗
FQI (Munos and Szepesvári, 2008)	✗	$O(\varepsilon^{-2})$	oracle-based	general	✓	✗
PSPI, practical (Xie et al., 2021)	✓	$O(\varepsilon^{-5})/O(\varepsilon^{-3})$	oracle-based	general / linear	✓	✗
PRO-RL (Zhan et al., 2022)	✓	$O(\varepsilon^{-6})$	oracle-based	general	✓	✗
ALMIS (Rashidinejad et al., 2023)	✓	$O(\varepsilon^{-2})$	oracle-based	general	✓	✗
A-CRAB (Zhu et al., 2023)	✓	$O(\varepsilon^{-2})$	oracle-based	general	✓	✗
PDOR (ours)	✓	$O(\varepsilon^{-4})$	$O(n)$	linear	✓	✓

Table 1: Comparison of selected methods for offline RL. The table shows some of the most relevant works for offline RL, and their characteristics. It is important to notice that many of these methods are designed for the general function approximation setting, while we focus on the easier setting of linear MDPs. However, most existing methods make use of oracles, which makes their computational complexity difficult to estimate, and while an efficient implementation can be derived by replacing the oracles appropriately, it is usually not immediate to prove sample complexity results for these practical versions.

π as $A_t \sim \pi(\cdot|X_t)$. Moreover, we define the following quantities for each policy π: its state-action value function $q^\pi(x,a) \doteq \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r(X_t,A_t) \mid X_0 = x, A_0 = a]$, its value function $v^\pi(x) \doteq \mathbb{E}_\pi[q^\pi(x,A_0)]$, its state occupancy measure $\nu^\pi(x) \doteq (1-\gamma)\mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t \mathbb{1}\{X_t = x\}]$, and its state-action occupancy measure $\mu^\pi(x,a) \doteq \pi(a|x)\nu^\pi(x)$. These quantities are known to satisfy the following useful relations, more commonly known respectively as Bellman's equation and flow constraint for policy π (Bellman, 1966):

$$\boldsymbol{q}^\pi = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v}^\pi \qquad \boldsymbol{\nu}^\pi = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{P}^\intercal \boldsymbol{\mu}^\pi. \quad (1)$$

Given this notation, we can also rewrite the normalized expected return in vector form as $\rho(\pi) = (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v}^\pi \rangle$ or equivalently as $\rho(\pi) = \langle \boldsymbol{r}, \boldsymbol{\mu}^\pi \rangle$.

Our work is based on the linear programming formulation due to Manne (1960b) (see also Puterman, 1994) which transforms the reinforcement learning problem into the search for an optimal state-action occupancy measure, obtained by solving the following Linear Program (LP):

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{r}, \boldsymbol{\mu} \rangle \\ \text{subject to} \quad & \boldsymbol{E}^\intercal \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{P}^\intercal \boldsymbol{\mu} \\ & \boldsymbol{\mu} \geq 0 \end{aligned} \qquad (2)$$

where $\boldsymbol{E} \in \mathbb{R}^{|\mathcal{X}\times\mathcal{A}|\times|\mathcal{X}|}$ denotes the matrix with components $\boldsymbol{E}_{(x,a),x'} \doteq \mathbb{1}\{x = x'\}$. The constraints of this LP are known to characterize the set of valid state-action occupancy measures. Therefore, an optimal solution $\boldsymbol{\mu}^*$ of the LP corresponds to the state-action occupancy measure associated to a policy π^* maximizing the expected return, and which is therefore optimal in the MDP. This policy can be extracted as $\pi^*(a|x) \doteq \mu^*(x,a) / \sum_{\bar{a}\in\mathcal{A}} \mu^*(x,\bar{a})$. However, this linear program cannot be directly solved in an efficient way in large MDPs due to the number of constraints and dimensions of the variables scaling with the size of the state space $\mathcal{X}$. Therefore, taking inspiration from the previous works of Bas-Serrano et al. (2021); Neu and Okolo (2023) we assume the knowledge of a *feature map* φ, which we then use to reduce the dimension of the problem. More specifically we consider the setting of Linear MDPs (Jin et al., 2020; Yang and Wang, 2019).

Definition 2.1 (Linear MDP)**.** An MDP is called linear if both the transition and reward functions can be expressed as a linear function of a given feature map $\varphi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$. That is, there exist $\psi : \mathcal{X} \to \mathbb{R}^d$ and $\boldsymbol{\omega} \in \mathbb{R}^d$ such that, for every $x, x' \in \mathcal{X}$ and $a \in \mathcal{A}$:

$$r(x,a) = \langle \boldsymbol{\varphi}(x,a), \boldsymbol{\omega} \rangle, \quad p(x' \mid x,a) = \langle \boldsymbol{\varphi}(x,a), \boldsymbol{\psi}(x') \rangle.$$

We assume that for all x, a, the norms of all relevant vectors are bounded by known constants as $\|\boldsymbol{\varphi}(x,a)\|_2 \leq D_{\boldsymbol{\varphi}}$, $\|\sum_{x'} \boldsymbol{\psi}(x')\|_2 \leq D_{\boldsymbol{\psi}}$, and $\|\boldsymbol{\omega}\|_2 \leq D_{\boldsymbol{\omega}}$. Moreover, we represent the feature map with the matrix $\boldsymbol{\Phi} \in \mathbb{R}^{|\mathcal{X}\times\mathcal{A}|\times d}$ with rows given by $\boldsymbol{\varphi}(x,a)^\intercal$, and similarly we define $\boldsymbol{\Psi} \in \mathbb{R}^{d\times|\mathcal{X}|}$ as the matrix with columns given by $\boldsymbol{\psi}(x)$.

With this notation we can rewrite the transition matrix as $\boldsymbol{P} = \boldsymbol{\Phi}\boldsymbol{\Psi}$. Furthermore, it is convenient to assume that the dimension d of the feature map cannot be trivially reduced, and therefore that the matrix $\boldsymbol{\Phi}$ is full-rank. An easily verifiable consequence of the Linear MDP assumption is that state-action value functions can be represented as a linear combinations of φ. That

is, there exist $\boldsymbol{\theta}^\pi \in \mathbb{R}^d$ such that:

$$\boldsymbol{q}^\pi = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{v}^\pi = \boldsymbol{\Phi}(\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}^\pi) = \boldsymbol{\Phi}\boldsymbol{\theta}^\pi. \tag{3}$$

It can be shown that for all policies π, the norm of $\boldsymbol{\theta}^\pi$ is at most $D_{\boldsymbol{\theta}} = D_{\boldsymbol{\omega}} + \frac{D_\Psi}{1-\gamma}$ (cf. Lemma B.1 in Jin et al., 2020). We then translate the linear program (2) to our setting, with the addition of the new variable $\boldsymbol{\lambda} \in \mathbb{R}^d$, resulting in the following new LP and its corresponding dual:

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{\omega}, \boldsymbol{\lambda} \rangle \\ \text{subject to} \quad & \boldsymbol{E}^\intercal \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^\intercal \boldsymbol{\lambda} \\ & \boldsymbol{\lambda} = \boldsymbol{\Phi}^\intercal \boldsymbol{\mu} \\ & \boldsymbol{\mu} \geq 0, \end{aligned} \tag{4}$$

$$\begin{aligned} \text{minimize} \quad & (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v} \rangle \\ \text{subject to} \quad & \boldsymbol{\theta} = \boldsymbol{\omega} + \gamma \boldsymbol{\Psi}\boldsymbol{v} \\ & \boldsymbol{E}\boldsymbol{v} \geq \boldsymbol{\Phi}\boldsymbol{\theta}. \end{aligned} \tag{5}$$

It can be immediately noticed how the introduction of $\boldsymbol{\lambda}$ did not change neither the set of admissible $\boldsymbol{\mu}$s nor the objective, and therefore did not alter the optimal solution. The Lagrangian associated to this set of linear programs is the function:

$$\begin{aligned} \mathfrak{L}(\boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v} \rangle + \langle \boldsymbol{\lambda}, \boldsymbol{\omega} + \gamma \boldsymbol{\Psi}\boldsymbol{v} - \boldsymbol{\theta} \rangle \\ &\quad + \langle \boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{E}\boldsymbol{v} \rangle \qquad (6) \\ &= \langle \boldsymbol{\lambda}, \boldsymbol{\omega} \rangle + \langle \boldsymbol{v}, (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^\intercal \boldsymbol{\lambda} - \boldsymbol{E}^\intercal \boldsymbol{\mu} \rangle \\ &\quad + \langle \boldsymbol{\theta}, \boldsymbol{\Phi}^\intercal \boldsymbol{\mu} - \boldsymbol{\lambda} \rangle. \qquad (7) \end{aligned}$$

It is known that finding optimal solutions $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ and $(\boldsymbol{v}^\star, \boldsymbol{\theta}^\star)$ for the primal and dual LPs is equivalent to finding a saddle point $(\boldsymbol{v}^\star, \boldsymbol{\theta}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$ of the Lagrangian function (Bertsekas, 1982). In the next section, we will develop primal-dual methods that aim to find approximate solutions to the above saddle-point problem, and convert these solutions to policies with near-optimality guarantees.

3 ALGORITHM AND MAIN RESULTS

This section introduces the concrete setting we study in this paper, and presents our main contributions.

We consider the offline-learning scenario where the agent has access to a dataset $\mathcal{D} = (W_t)_{t=1}^n$, collected by a behavior policy π_B, and composed of n random observations of the form $W_t = (X_t^0, X_t, A_t, R_t, X_t')$. The random variables X_t^0, (X_t, A_t) and X_t' are sampled, respectively, from the initial-state distribution ν_0, the discounted occupancy measure of the behavior policy, denoted as μ_B, and from $p(\cdot \mid X_t, A_t)$. Finally, R_t denotes the reward $r(X_t, A_t)$. We assume that all observations W_t are generated independently of each other, and will often use the notation $\boldsymbol{\varphi}_t = \boldsymbol{\varphi}(X_t, A_t)$.

Our strategy consists in finding approximately good solutions for the LPs (4) and (5) using stochastic optimization methods, which require access to unbiased gradient estimates of the Lagrangian (Equation 7). The main challenge we need to overcome is constructing suitable estimators based only on observations drawn from the behavior policy. We address this challenge by introducing the matrix $\boldsymbol{\Lambda} = \mathbb{E}_{X,A\sim\mu_B}[\boldsymbol{\varphi}(X,A)\boldsymbol{\varphi}(X,A)^\intercal]$ (supposed to be invertible for the sake of argument for now), and rewriting the gradient with respect to $\boldsymbol{\lambda}$ as

$$\begin{aligned} \nabla_{\boldsymbol{\lambda}} \mathfrak{L}(\boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{v}, \boldsymbol{\theta}) &= \boldsymbol{\omega} + \gamma \boldsymbol{\Psi}\boldsymbol{v} - \boldsymbol{\theta} \\ &= \boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\left(\boldsymbol{\omega} + \gamma \boldsymbol{\Psi}\boldsymbol{v} - \boldsymbol{\theta}\right) \\ &= \boldsymbol{\Lambda}^{-1}\mathbb{E}\left[\boldsymbol{\varphi}(X_t, A_t)\boldsymbol{\varphi}(X_t, A_t)^\intercal \left(\boldsymbol{\omega} + \gamma \boldsymbol{\Psi}\boldsymbol{v} - \boldsymbol{\theta}\right)\right] \\ &= \boldsymbol{\Lambda}^{-1}\mathbb{E}\left[\boldsymbol{\varphi}(X_t, A_t)\left(R_t + \gamma \boldsymbol{v}(X_t') - \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(X_t, A_t)\rangle\right)\right]. \end{aligned}$$

This suggests that the vector within the expectation can be used to build an unbiased estimator of the desired gradient. A downside of using this estimator is that it requires knowledge of $\boldsymbol{\Lambda}$. However, this can be sidestepped by a reparametrization trick inspired by Nachum and Dai (2020): introducing the parametrization $\boldsymbol{\beta} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\lambda}$, the objective can be rewritten as

$$\begin{aligned} \mathfrak{L}(\boldsymbol{\beta}, \boldsymbol{\mu}; \boldsymbol{v}, \boldsymbol{\theta}) = (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v} \rangle + \langle \boldsymbol{\beta}, \boldsymbol{\Lambda}\big(\boldsymbol{\omega} + \gamma \boldsymbol{\Psi}\boldsymbol{v} - \boldsymbol{\theta}\big) \rangle \\ + \langle \boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{E}\boldsymbol{v} \rangle. \end{aligned}$$

This can be indeed seen to generalize the tabular reparametrization of Nachum and Dai (2020) to the case of linear function approximation. Notably, our linear reparametrization does not change the structure of the saddle-point problem, but allows building an unbiased estimator of $\nabla_{\boldsymbol{\beta}} \mathfrak{L}(\boldsymbol{\beta}, \boldsymbol{\mu}; \boldsymbol{v}, \boldsymbol{\theta})$ without knowledge of $\boldsymbol{\Lambda}$ as

$$\tilde{\boldsymbol{g}}_{\boldsymbol{\beta}} = \boldsymbol{\varphi}(X_t, A_t)\left(R_t + \gamma \boldsymbol{v}(X_t') - \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(X_t, A_t)\rangle\right).$$

In what follows, we will use the more general parametrization $\boldsymbol{\beta} = \Lambda^{-c}\lambda$, with $c \in \{1/2, 1\}$, and construct a primal-dual stochastic optimization method that can be implemented efficiently in the offline setting based on the observations above. Using $c = 1$ allows to run our algorithm without knowledge of $\boldsymbol{\Lambda}$, that is, without knowing the behavior policy that generated the dataset, while using $c = 1/2$ results in a tighter bound[1], at the price of having to assume knowledge of $\boldsymbol{\Lambda}$.

Our algorithm (presented as Algorithm 1) is inspired by the method of Neu and Okolo (2023), originally

[1] By "tighter bound" we refer to dependence on the coverage ratio introduced in Definition 3.1. We give more details on this in Section 6.

designed for planning with a generative model. The algorithm has a double-loop structure, where at each iteration t we run one step of stochastic gradient ascent for $\boldsymbol{\beta}$, and also an inner loop which runs K iterations of stochastic gradient descent on $\boldsymbol{\theta}$ making sure that $\langle \boldsymbol{\varphi}(x,a), \boldsymbol{\theta}_t \rangle$ is a good approximation of the true action-value function of π_t. Iterations of the inner loop are indexed by k. The main idea of the algorithm is to compute the unbiased estimators $\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,k}$ and $\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t}$ of the gradients $\nabla_{\boldsymbol{\theta}} \mathfrak{L}(\boldsymbol{\beta}_t, \boldsymbol{\mu}_t; \cdot, \boldsymbol{\theta}_{t,k})$ and $\nabla_{\boldsymbol{\beta}} \mathfrak{L}(\boldsymbol{\beta}_t, \cdot; \boldsymbol{v}_t, \boldsymbol{\theta}_t)$, and use them to update the respective variables iteratively. We then define a softmax policy π_t at each iteration t using the $\boldsymbol{\theta}$ parameters as $\pi_t(a|x) = \sigma\left(\alpha \sum_{i=1}^{t-1} \langle \boldsymbol{\varphi}(x,a), \boldsymbol{\theta}_i \rangle\right)$. The other higher-dimensional variables $(\boldsymbol{\mu}_t, \boldsymbol{v}_t)$ are defined symbolically in terms of $\boldsymbol{\beta}_t$, $\boldsymbol{\theta}_t$ and π_t, and used only as auxiliary variables for computing the estimates $\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,k}$ and $\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t}$. Specifically, we set these variables as

$$v_t(x) = \sum_a \pi_t(a|x) \langle \boldsymbol{\varphi}(x,a), \boldsymbol{\theta}_t \rangle, \tag{8}$$

$$\mu_{t,k}(x,a) = \pi_t(a|x)\big((1-\gamma)\mathbb{1}\{X^0_{t,k} = x\} + \gamma \langle \boldsymbol{\varphi}_{t,k}, \boldsymbol{\Lambda}^{c-1} \boldsymbol{\beta}_t \rangle \mathbb{1}\{X'_{t,k} = x\}\big). \tag{9}$$

Finally, the gradient estimates can be defined as

$$\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t} = \boldsymbol{\Lambda}^{c-1} \boldsymbol{\varphi}_t \left(R_t + \gamma v_t(X'_t) - \langle \boldsymbol{\varphi}_t, \boldsymbol{\theta}_t \rangle\right), \tag{10}$$

$$\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,k} = \boldsymbol{\Phi}^\top \boldsymbol{\mu}_{t,k} - \boldsymbol{\Lambda}^{c-1} \boldsymbol{\varphi}_{t,k} \langle \boldsymbol{\varphi}_{t,k}, \boldsymbol{\beta}_t \rangle. \tag{11}$$

These gradient estimates are then used in a projected gradient ascent/descent scheme, with the ℓ_2 projection operator denoted by Π. The feasible sets of the two parameter vectors are chosen as ℓ_2 balls of radii D_θ and D_β, denoted respectively as $\mathbb{B}(D_\theta)$ and $\mathbb{B}(D_\beta)$. Notably, the algorithm does not need to compute $v_t(x)$, $\mu_{t,k}(x,a)$, or $\pi_t(a|x)$ for all states x, but only for the states that are accessed during the execution of the method. In particular, π_t does not need to be computed explicitly, and it can be efficiently represented by the single d-dimensional parameter vector $\sum_{i=1}^t \boldsymbol{\theta}_i$.

Due to the double-loop structure, each iteration t uses K samples from the dataset $\mathcal{D}$, adding up to a total of $n = KT$ samples over the course of T iterations. Each gradient update calculated by the method uses a constant number of elementary vector operations, resulting in a total computational complexity of $O(|\mathcal{A}|dn)$ elementary operations. At the end, our algorithm outputs a policy selected uniformly at random from the T iterations.

3.1 Main result

We are now almost ready to state our main result. Before doing so, we first need to discuss the quantities appearing in the guarantee, and provide an intuitive explanation for them.

Algorithm 1 Primal-Dual Offline RL (PDOR)

Input: Learning rates α, ζ, η, initial points $\boldsymbol{\theta}_0 \in \mathbb{B}(D_\theta), \boldsymbol{\beta}_1 \in \mathbb{B}(D_\beta), \pi_1$, and data $\mathcal{D} = (W_t)_{t=1}^n$
for $t = 1$ **to** T **do**
  Initialize $\boldsymbol{\theta}_{t,1} = \boldsymbol{\theta}_{t-1}$
  for $k = 1$ **to** $K-1$ **do**
    Obtain sample $W_{t,k} = (X^0_{t,k}, X_{t,k}, A_{t,k}, X'_{t,k})$
    $\boldsymbol{\mu}_{t,k} = \pi_t \circ \big[(1-\gamma)\boldsymbol{e}_{X^0_{t,k}} + \gamma \langle \boldsymbol{\varphi}(X_{t,k}, A_{t,k}), \boldsymbol{\Lambda}^{c-1} \boldsymbol{\beta}_t \rangle \boldsymbol{e}_{X'_{t,k}}\big]$
    $\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,i} = \boldsymbol{\Phi}^\top \boldsymbol{\mu}_{t,k} - \boldsymbol{\Lambda}^{c-1} \boldsymbol{\varphi}(X_{t,k}, A_{t,k}) \langle \boldsymbol{\varphi}(X_{t,k}, A_{t,k}), \boldsymbol{\beta}_t \rangle$
    $\boldsymbol{\theta}_{t,k+1} = \Pi_{\mathbb{B}(D_\theta)}(\boldsymbol{\theta}_{t,k} - \eta \tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,i})$ *// Stochastic gradient descent*
  end for
  $\boldsymbol{\theta}_t = \frac{1}{K} \sum_{k=1}^K \boldsymbol{\theta}_{t,k}$

  Obtain sample $W_t = (X^0_t, X_t, A_t, X'_t)$
  $\boldsymbol{v}_t = \boldsymbol{E}^\top \big(\pi_t \circ \boldsymbol{\Phi} \boldsymbol{\theta}_t\big)$
  $\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t} = \boldsymbol{\Lambda}^{c-1} \boldsymbol{\varphi}(X_t, A_t)\big(R_t + \gamma \boldsymbol{v}_t(X'_t) - \langle \boldsymbol{\varphi}(X_t, A_t), \boldsymbol{\theta}_t \rangle\big)$
  $\boldsymbol{\beta}_{t+1} = \Pi_{\mathbb{B}(D_\beta)}(\boldsymbol{\beta}_t + \zeta \tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t})$ *// Stochastic gradient ascent*

  $\pi_{t+1} = \sigma(\alpha \sum_{i=1}^t \boldsymbol{\Phi} \boldsymbol{\theta}_i)$ *// Policy update*
end for
return π_J with $J \sim \mathcal{U}(T)$.

Similarly to previous work, we capture the partial coverage assumption by expressing the rate of convergence to the optimal policy in terms of a *coverage ratio* that measures the mismatch between the behavior and the optimal policy. Several definitions of coverage ratio are surveyed by Uehara and Sun (2022). In this work, we employ a notion of *feature* coverage ratio for linear MDPs that defines coverage in feature space rather than in state-action space, similarly to Jin et al. (2021), but with a smaller ratio.

Definition 3.1. Let $c \in \{1/2, 1\}$. For a policy π, we denote by $\overline{\boldsymbol{\varphi}}(\pi) = \mathbb{E}_{X,A \sim \mu^\pi}[\boldsymbol{\varphi}(X,A)]$ the average feature vector under π. We define the generalized coverage ratio as[2]

$$C_{\varphi,c}(\pi^*; \pi_B) = \overline{\boldsymbol{\varphi}}(\pi^*)^\top \boldsymbol{\Lambda}^{-2c} \overline{\boldsymbol{\varphi}}(\pi^*).$$

We defer a detailed discussion of this ratio to Section 6, where we compare it with similar notions in the literature. We are now ready to state our main result.

Theorem 3.2. *Given a linear MDP (Definition 2.1) such that $\boldsymbol{\theta}^\pi \in \mathbb{B}(D_{\boldsymbol{\theta}})$ for any policy π. Assume that the coverage ratio is bounded $C_{\varphi,c}(\pi^*; \pi_B) \le D^2_{\boldsymbol{\beta}}$. Then,*

[2]When $\boldsymbol{\Lambda}$ is not invertible but $\overline{\boldsymbol{\varphi}}(\pi^*)$ is in the column space of $\boldsymbol{\Lambda}$, we can define the coverage ratio using the Moore-Penrose pseudoinverse, and set it to $+\infty$ otherwise.

for any comparator policy π^*, the policy output by an appropriately tuned instance of Algorithm 1 satisfies $\mathbb{E}\left[\langle \boldsymbol{\mu}^{\pi^*} - \boldsymbol{\mu}^{\boldsymbol{\pi}_{out}}, \boldsymbol{r}\rangle\right] \le \varepsilon$ with a number of samples n_ϵ that is $O\left(\varepsilon^{-4} D_{\boldsymbol{\theta}}^4 D_{\boldsymbol{\varphi}}^{8c} D_{\boldsymbol{\beta}}^4 d^{2-2c} \log|\mathcal{A}|\right)$.

The concrete parameter choices are detailed in the full version of the theorem in Appendix A. The main theorem can be simplified by making some standard assumptions, formalized by the following corollary.

Corollary 3.3. *Assume that the bound of the feature vectors $D_{\boldsymbol{\varphi}}$ is of order $O(1)$, that $D_{\boldsymbol{\omega}} = D_{\boldsymbol{\psi}} = \sqrt{d}$ and that $D_{\boldsymbol{\beta}}^2 = c \cdot C_{\varphi,c}(\pi^*; \pi_B)$ for some positive universal constant c. Then, under the same assumptions of Theorem 3.2, n_ε is of order $O\left(\frac{d^4 C_{\varphi,c}(\pi^*;\pi_B)^2 \log|\mathcal{A}|}{d^{2c}(1-\gamma)^4 \varepsilon^4}\right)$.*

4 ANALYSIS

This section explains the rationale behind some of the technical choices of our algorithm, and sketches the proof of our main result.

First, we explicitly rewrite the expression of the Lagrangian (7), after performing the change of variable $\boldsymbol{\lambda} = \boldsymbol{\Lambda}^c \boldsymbol{\beta}$:

$$\begin{aligned}\mathfrak{L}(\boldsymbol{\beta}, \boldsymbol{\mu}; \boldsymbol{v}, \boldsymbol{\theta}) &= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v}\rangle + \langle \boldsymbol{\beta}, \boldsymbol{\Lambda}^c\big(\boldsymbol{\omega} + \gamma \boldsymbol{\Psi} \boldsymbol{v} - \boldsymbol{\theta}\big)\rangle \\ &\quad + \langle \boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{E}\boldsymbol{v}\rangle \qquad (12) \\ &= \langle \boldsymbol{\beta}, \boldsymbol{\Lambda}^c \boldsymbol{\omega}\rangle + \langle \boldsymbol{v}, (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^\intercal \boldsymbol{\Lambda}^c \boldsymbol{\beta} - \boldsymbol{E}^\intercal \boldsymbol{\mu}\rangle \\ &\quad + \langle \boldsymbol{\theta}, \boldsymbol{\Phi}^\intercal \boldsymbol{\mu} - \boldsymbol{\Lambda}^c \boldsymbol{\beta}\rangle. \qquad (13)\end{aligned}$$

We aim to find an approximate saddle-point of the above convex-concave objective function. One challenge that we need to face is that the variables $\boldsymbol{v}$ and $\boldsymbol{\mu}$ have dimension proportional to the size of the state space $|\mathcal{X}|$, so making explicit updates to these parameters would be prohibitively expensive in MDPs with large state spaces. To address this challenge, we choose to parametrize $\boldsymbol{\mu}$ in terms of a policy π and $\boldsymbol{\beta}$ through the symbolic assignment $\boldsymbol{\mu} = \boldsymbol{\mu}_{\boldsymbol{\beta},\pi}$, where

$$\mu_{\boldsymbol{\beta},\pi}(x,a) \doteq \pi(a|x)\Big[(1-\gamma)\nu_0(x) + \gamma\langle \boldsymbol{\psi}(x), \boldsymbol{\Lambda}^c \boldsymbol{\beta}\rangle\Big].$$

This choice can be seen to satisfy the first constraint of the primal LP (4), and thus the gradient of the Lagrangian (13) evaluated at $\boldsymbol{\mu}_{\boldsymbol{\beta},\pi}$ with respect to $\boldsymbol{v}$ can be verified to be 0. This parametrization makes it possible to express the Lagrangian as a function of only $\boldsymbol{\theta}, \boldsymbol{\beta}$ and π as

$$\begin{aligned}f(\boldsymbol{\theta}, \boldsymbol{\beta}, \pi) &\doteq \mathfrak{L}(\boldsymbol{\beta}, \boldsymbol{\mu}_{\boldsymbol{\beta},\pi}; \boldsymbol{v}, \boldsymbol{\theta}) \\ &= \langle \boldsymbol{\beta}, \boldsymbol{\Lambda}^c \boldsymbol{\omega}\rangle + \langle \boldsymbol{\theta}, \boldsymbol{\Phi}^\intercal \boldsymbol{\mu}_{\boldsymbol{\beta},\pi} - \boldsymbol{\Lambda}^c \boldsymbol{\beta}\rangle. \qquad (14)\end{aligned}$$

For convenience, we also define the quantities $\boldsymbol{\nu}_{\boldsymbol{\beta}} = \boldsymbol{E}^\intercal \boldsymbol{\mu}_{\boldsymbol{\beta},\pi}$ and $v_{\boldsymbol{\theta},\pi}(s) \doteq \sum_a \pi(a|s)\,\langle \boldsymbol{\theta}, \boldsymbol{\varphi}(x,a)\rangle$, which enables us to rewrite f as

$$\begin{aligned}f(\boldsymbol{\theta}, \boldsymbol{\beta}, \pi) &= \langle \boldsymbol{\Lambda}^c \boldsymbol{\beta}, \boldsymbol{\omega} - \boldsymbol{\theta}\rangle + \langle \boldsymbol{v}_{\boldsymbol{\theta},\pi}, \boldsymbol{\nu}_{\boldsymbol{\beta}}\rangle \\ &= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v}_{\boldsymbol{\theta},\pi}\rangle \\ &\quad + \langle \boldsymbol{\Lambda}^c \boldsymbol{\beta}, \boldsymbol{\omega} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\theta},\pi} - \boldsymbol{\theta}\rangle. \qquad (15)\end{aligned}$$

The above choices allow us to perform stochastic gradient / ascent over the low-dimensional parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ and the policy π. In order to calculate an unbiased estimator of the gradients, we first observe that the choice of $\mu_{t,k}$ in Algorithm 1 is an unbiased estimator of $\mu_{\boldsymbol{\beta}_t,\pi_t}$:

$$\begin{aligned}\mathbb{E}_{t,k}\left[\mu_{t,k}(x,a)\right] &= \pi_t(a|x)\Big((1-\gamma)\mathbb{P}(X_{t,k}^0 = x) \\ &\qquad + \mathbb{E}_{t,k}\left[\mathbb{1}\{X'_{t,k} = x\}\langle \boldsymbol{\varphi}_t, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\rangle\right]\Big) \\ &= \pi_t(a|x)\Big((1-\gamma)\nu_0(x) \\ &\qquad + \gamma \sum_{\bar{x},\bar{a}} \mu_B(\bar{x},\bar{a}) p(x|\bar{x},\bar{a}) \boldsymbol{\varphi}(\bar{x},\bar{a})^\intercal \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\Big) \\ &= \pi_t(a|x)\Big((1-\gamma)\nu_0(x) + \gamma \boldsymbol{\psi}(x)^\intercal \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\Big) \\ &= \mu_{\boldsymbol{\beta}_t,\pi_t}(x,a),\end{aligned}$$

where we used the fact that $p(x|\bar{x},\bar{a}) = \langle \boldsymbol{\psi}(x), \boldsymbol{\varphi}(\bar{x},\bar{a})\rangle$, and the definition of $\boldsymbol{\Lambda}$. This in turn facilitates proving that the gradient estimate $\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,k}$, defined in Equation 11, is indeed unbiased:

$$\begin{aligned}\mathbb{E}_{t,k}\left[\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,k}\right] &= \boldsymbol{\Phi}^\intercal \mathbb{E}_{t,k}\left[\boldsymbol{\mu}_{t,k}\right] - \boldsymbol{\Lambda}^{c-1}\mathbb{E}_{t,k}\left[\boldsymbol{\varphi}_{t,k}\boldsymbol{\varphi}_{t,k}^\intercal\right]\boldsymbol{\beta}_t \\ &= \boldsymbol{\Phi}^\intercal \boldsymbol{\mu}_{\boldsymbol{\beta}_t,\pi_t} - \boldsymbol{\Lambda}^c \boldsymbol{\beta}_t = \nabla_{\boldsymbol{\theta}} \mathfrak{L}(\boldsymbol{\beta}_t, \boldsymbol{\mu}_t; \boldsymbol{v}_t, \cdot).\end{aligned}$$

A similar proof is used for $\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t}$ and is detailed in Appendix B.3.

Our analysis is based on arguments by Neu and Okolo (2023), carefully adapted to the reparametrized version of the Lagrangian presented above. The proof studies the following central quantity that we refer to as *dynamic duality gap*:

$$\mathcal{G}_T(\boldsymbol{\beta}^*, \pi^*; \boldsymbol{\theta}_{1:T}^*) \doteq \frac{1}{T}\sum_{t=1}^T (f(\boldsymbol{\beta}^*, \pi^*; \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}_t, \pi_t; \boldsymbol{\theta}_t^*)).$$

Here, $(\boldsymbol{\theta}_t, \boldsymbol{\beta}_t, \pi_t)$ are the iterates of the algorithm, $\boldsymbol{\theta}_{1:T}^* = (\boldsymbol{\theta}_t^*)_{t=1}^T$ a sequence of comparators for $\boldsymbol{\theta}$, and finally $\boldsymbol{\beta}^*$ and π^* are fixed comparators for $\boldsymbol{\beta}$ and π, respectively. Our first key lemma relates the suboptimality of the output policy to $\mathcal{G}_T$ for a specific choice of comparators.

Lemma 4.1. *Let $\boldsymbol{\theta}_t^* \doteq \boldsymbol{\theta}^{\pi_t}$, π^* be any policy, and $\beta^* = \boldsymbol{\Lambda}^{-c}\boldsymbol{\Phi}^\top \boldsymbol{\mu}^{\pi^*}$. Then, $\mathbb{E}\left[\langle \boldsymbol{\mu}^{\pi^*} - \boldsymbol{\mu}^{\boldsymbol{\pi}_{out}}, \boldsymbol{r}\rangle\right] = \mathcal{G}_T\big(\boldsymbol{\beta}^*, \pi^*; \boldsymbol{\theta}_{1:T}^*\big)$.*

The proof is relegated to Appendix B.1. Our second key lemma rewrites the gap $\mathcal{G}_T$ for *any* choice of comparators as the sum of three regret terms:

Lemma 4.2. *With the choice of comparators of Lemma 4.1*

$$\mathcal{G}_T(\boldsymbol{\beta}^*, \pi^*; \boldsymbol{\theta}^*_{1:T}) = \frac{1}{T}\sum_{t=1}^{T}\Big(\langle \boldsymbol{\theta}_t - \boldsymbol{\theta}^*_t, g_{\boldsymbol{\theta},t}\rangle + \langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_t, g_{\boldsymbol{\beta},t}\rangle + \sum_s \nu^{\pi^*}(s)\sum_a (\pi^*(a|s) - \pi_t(a|s))\langle \boldsymbol{\theta}_t, \boldsymbol{\varphi}(x,a)\rangle\Big),$$

where $g_{\boldsymbol{\theta},t} = \boldsymbol{\Phi}^\top \boldsymbol{\mu}_{\boldsymbol{\beta}_t,\pi_t} - \boldsymbol{\Lambda}^c \boldsymbol{\beta}_t$ *and* $g_{\boldsymbol{\beta},t} = \boldsymbol{\Lambda}^c(\boldsymbol{\omega} + \gamma \boldsymbol{\Psi} v_{\boldsymbol{\theta}_t,\pi_t} - \boldsymbol{\theta}_t)$.

The proof is presented in Appendix B.2. To conclude the proof we bound the three terms appearing in Lemma 4.2. The first two of those are bounded using standard gradient descent/ascent analysis (Lemmas B.1 and B.2), while for the latter we use mirror descent analysis (Lemma B.3). The details of these steps are reported in Appendix B.3.

5 EXTENSION TO AVERAGE-REWARD MDPS

In this section, we briefly explain how to extend our approach to offline learning in *average reward MDPs*, establishing the first sample complexity result for this setting. After introducing the setup, we outline a remarkably simple adaptation of our algorithm along with its performance guarantees for this setting. The reader is referred to Appendix C for the full details, and to Chapter 8 of Puterman (1994) for a more thorough discussion of average-reward MDPs.

In the average reward setting we aim to optimize the objective $\rho^\pi(x) = \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi\big[\sum_{t=1}^T r(x_t, a_t) \;\big|\; x_1 = x\big]$, representing the long-term average reward of policy π when started from state $x \in \mathcal{X}$. Unlike the discounted setting, the average reward criterion prioritizes long-term frequency over proximity of good rewards due to the absence of discounting which expresses a preference for earlier rewards. As is standard in the related literature, we will assume that ρ^π is well-defined for any policy and is independent of the start state, and thus will use the same notation to represent the scalar average reward of policy π. Due to the boundedness of the rewards, we clearly have $\rho^\pi \in [0,1]$. Similarly to the discounted setting, it is possible to define quantities analogous to the value and action value functions as the solutions to the Bellman equations $\boldsymbol{q}^\pi = \boldsymbol{r} - \rho^\pi \mathbf{1} + \boldsymbol{P}\boldsymbol{v}^\pi$, where $\boldsymbol{v}^\pi$ is related to the action-value function as $v^\pi(x) = \sum_a \pi(a|x) q^\pi(x,a)$. We will make the following standard assumption about the MDP (see Section 17.4 of Meyn and Tweedie (1996)):

Assumption 5.1. For all stationary policies π, the Bellman equations have a solution $\boldsymbol{q}^\pi$ satisfying $\sup_{x,a} q^\pi(x,a) - \inf_{x,a} q^\pi(x,a) < D_q$.

Furthermore, we will continue to work with the linear MDP assumption of Definition 2.1, and will additionally make the following minor assumption:

Assumption 5.2. The all ones vector $\mathbf{1}$ is contained in the column span of the feature matrix $\boldsymbol{\Phi}$. Furthermore, let $\boldsymbol{\varrho} \in \mathbb{R}^d$ such that for all $(x,a) \in \mathcal{Z}$, $\langle \boldsymbol{\varphi}(x,a), \boldsymbol{\varrho}\rangle = 1$.

Using these insights, it is straightforward to derive a linear program akin to (2) that characterize the optimal occupancy measure and thus an optimal policy in average-reward MDPs. Starting from this formulation and proceeding as in Sections 2 and 4, we equivalently restate this optimization problem as finding the saddle-point of the reparametrized Lagrangian defined as follows:

$$\mathfrak{L}(\boldsymbol{\beta}, \boldsymbol{\mu}; \rho, \boldsymbol{v}, \boldsymbol{\theta}) = \rho + \langle \boldsymbol{\beta}\,, \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v} - \boldsymbol{\theta} - \rho\boldsymbol{\varrho}]\rangle + \langle \boldsymbol{\mu}\,, \boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{E}\boldsymbol{v}\rangle.$$

As previously, the saddle point can be shown to be equivalent to an optimal occupancy measure under the assumption that the MDP is linear in the sense of Definition 2.1. Notice that the above Lagrangian slightly differs from that of the discounted setting in Equation (12) due to the additional optimization parameter ρ, but otherwise our main algorithm can be directly generalized to this objective. We present details of the derivations and the resulting algorithm in Appendix C. The following theorem states the performance guarantees for this method.

Theorem 5.3. *Given a linear MDP (Definition 2.1) satisfying Assumption 5.2 and such that* $\boldsymbol{\theta}^\pi \in \mathbb{B}(D_{\boldsymbol{\theta}})$ *for any policy* π. *Assume that the coverage ratio is bounded* $C_{\varphi,c}(\pi^*; \pi_B) \le D^2_{\boldsymbol{\beta}}$. *Then, for any comparator policy* π^*, *the policy output by an appropriately tuned instance of Algorithm 2 satisfies* $\mathbb{E}\left[\langle \boldsymbol{\mu}^{\pi^*} - \boldsymbol{\mu}^{\boldsymbol{\pi}_{out}}, \boldsymbol{r}\rangle\right] \le \varepsilon$ *with a number of samples* n_ϵ *that is* $O\left(\varepsilon^{-4} D^4_{\boldsymbol{\theta}} D^{12c-2}_{\boldsymbol{\varphi}} D^4_{\boldsymbol{\beta}} d^{2-2c} \log|\mathcal{A}|\right)$.

As compared to the discounted case, this additional dependence of the sample complexity on $D_{\boldsymbol{\varphi}}$ is due to the extra optimization variable ρ. We provide the full proof of this theorem along with further discussion in Appendix C.

6 DISCUSSION AND FINAL REMARKS

In this section, we compare our results with the most relevant ones from the literature, with a particular focus on discussing the relations between the coverage ratios used in our work and the ones used in related literature. Our Table 1 can be used as a reference. As a complement to this section, we refer the interested

reader to the recent work by Uehara and Sun (2022), which provides a survey of offline RL methods with their coverage and structural assumptions. Detailed computations can be found in Appendix E.

An important property of our method is that it only requires partial coverage. This sets it apart from classic batch RL methods like fitted Q-iteration (Ernst et al., 2005; Munos and Szepesvári, 2008; Chen and Jiang, 2019), whose analysis requires a stronger uniform-coverage assumption. Interestingly, our results defy the common wisdom in the related literature that suggests that obtaining guarantees under weaker partial-coverage assumptions requires the use of pessimistic adjustments (e.g., Jin et al. (2021); Xie et al. (2021))—indeed, notice that our algorithm does not implement any form of explicit pessimism. In fact, as we argue below, the notion of coverage that our guarantees depend on is in many senses much weaker than the most commonly used notions appearing in the literature.

Let us review some existing notions of coverage and contrast them to our notion. Jin et al. (2021) (Theorem 4.4) rely on a *feature* coverage ratio which can be written as

$$C^{\circ}(\pi^*;\pi_B) = \mathbb{E}_{X,A\sim\mu^*}\left[\sqrt{\boldsymbol{\varphi}(X,A)^{\intercal}\boldsymbol{\Lambda}^{-1}\boldsymbol{\varphi}(X,A)}\right]. \tag{16}$$

Although, in general, $C_{\varphi,c}$ is incomparable with C°, a simple geometric argument shows the advantages of our coverage. A boundedness condition on $C^{\circ}(\pi^*;\pi_B)$ requires the column space of $\boldsymbol{\Lambda}$ to span the subspace of $\mathbb{R}^d$ spanned by optimal features, $\operatorname{span}\{\boldsymbol{\varphi}(x,a)|\mu^{\pi^*}(x,a) > 0\}$. In contrast, boundedness of $C_{\varphi,c}$ only requires $\overline{\boldsymbol{\varphi}}(\pi^*) \in \operatorname{range}(\boldsymbol{\Lambda})$. Intuitively, we only require the behavior policy to witness a single direction in feature space (the average feature vector under π^*) compared to a whole, potentially d-dimensional, subspace. This can make a big difference, especially when d is large. In Appendix E, we show an example where C° can be arbitrarily larger than both $C_{\varphi,1/2}$ and $C_{\varphi,1}$.

This kind of coverage ratio has appeared in the literature before, but only for finite-horizon problems. Concretely, Zanette et al. (2021) propose a computationally intense algorithm that demonstrates a regret bound scaling with a quantity essentially equivalent to our $C_{\varphi,1/2}$. Uehara and Sun (2022) and Zhang et al. (2022) use a coverage ratio that is conceptually similar to Equation (16),

$$C^{\dagger}(\pi^*;\pi_B) = \sup_{y\in\mathbb{R}^d} \frac{y^{\intercal}\boldsymbol{\Lambda}^* y}{y^{\intercal}\boldsymbol{\Lambda} y}, \tag{17}$$

where $\boldsymbol{\Lambda}^* = \mathbb{E}_{X,A\sim\mu^*}[\boldsymbol{\varphi}(X,A)\boldsymbol{\varphi}(X,A)^{\intercal}]$. Some linear algebra shows that $C_{\varphi,1/2} \leq dC^{\dagger}$. It should be noted that the algorithm from Uehara and Sun (2022) also works with unknown features, at the cost of being computationally inefficient. The algorithm from Zhang et al. (2022) instead is limited to the finite-horizon setting.

We can gain some further insight from the special case of tabular MDPs, although it is hard to compare our ratio with existing ones there, because in this setting, error bounds are commonly stated in terms of $\sup_{x,a} {}^{\mu^*(x,a)}/_{\mu_B(x,a)}$, often introducing an explicit dependency on the number of states (e.g., Liu et al., 2020). However, looking at how the coverage ratio specializes to the tabular setting can still provide some insight. First, $C_{\varphi,1/2}(\pi^*;\pi_B) = \sum_{x,a} {}^{(\mu^*(x,a))^2}/_{\mu_B(x,a)}$, which of course is smaller than the more standard $C^{\diamond}(\pi^*;\pi_B) = \sum_{x,a} {}^{\mu^*(x,a)}/_{\mu_B(x,a)}$. Interestingly, $C_{\varphi,1/2}(\pi^*;\pi_B) = 1 + \mathcal{X}^2(\mu^*\|\mu_B)$, where $\mathcal{X}^2$ denotes the chi-square divergence, a crucial quantity in off-distribution learning based on importance sampling (Cortes et al., 2010). An analogous quantity was used by Li et al. (2014) to characterize the sample complexity of off-policy policy evaluation. Unfortunately, $C_{\varphi,1}(\pi^*;\pi_B) = \sum_{x,a} ({}^{\mu^*(x,a)}/_{\mu_B(x,a)})^2$ is non-comparable with $C^{\diamond}$ in general, and larger than $C_{\varphi,1/2}$. A similar quantity to $C_{\varphi,1}$ was used by Lykouris et al. (2021) in the context of RL with adversarial corruptions.

The most directly comparable works to ours are those of Xie et al. (2021) and Cheng et al. (2022), which are the only known practical methods to consider function approximation in the infinite-horizon setting, with minimal assumptions on the dataset. They both use the coverage ratio $C_{\mathcal{F}}(\pi^*;\pi_B) = \max_{f\in\mathcal{F}} {}^{\|f-\mathcal{T}f\|^2_{\mu^*}}/_{\|f-\mathcal{T}f\|^2_{\mu_B}}$, where $\mathcal{F}$ is a function class and $\mathcal{T}$ the Bellman operator. This can be shown to reduce to Equation (17) for linear MDPs (cf. Appendix E). However, the specialized bound of Xie et al. (2021) (Theorem 3.2) scales with the potentially larger ratio from Equation (16). Both their algorithms have superlinear computational complexity and a sample complexity of $O(\varepsilon^{-5})$. While the authors make plausible arguments in their paper that their method can be efficiently implemented in the linear setting and may obtain a sample complexity of order of order ε^{-2}, these statements are not supported with rigorous proofs. Hence, our result is technically the first *provably* computationally effective method that achieves a rate better than $O(\varepsilon^{-5})$, with the additional benefit of using a single-direction coverage ratio as discussed in the above paragraphs.

The above discussion outlines two major open problems that we leave open for future work. First, we highlight that so far, no computationally efficient algorithm exists for our setting that achieves the minimax optimal sample complexity rate of $O(\varepsilon^{-2})$ (Xiao et al., 2021; Rashidinejad et al., 2022). Regarding our own algorithm, it is clear that the extra $O(\varepsilon^{-2})$ factor in

our bounds is due to the nested-loop structure of the algorithm. How to remove this component from our algorithm design is currently unclear, but we suspect that that borrowing ideas from the literature on optimistic descent methods (Korpelevich, 1976; Rakhlin and Sridharan, 2013) or two-timescale stochastic approximation (Borkar, 1997) may bring us closer to an answer. A second limitation of our contribution is that, in order to scale with $C_{\varphi,1/2}$, our method requires prior knowledge of $\mathbf{\Lambda}$. We believe that this limitation can be relaxed at the price of a significantly more involved analysis, for instance by setting aside some fraction of the data set to estimate $\mathbf{\Lambda}$ (or directly $\mathbf{\Lambda}^{-1}$, using techniques from (Neu and Olkhovskaya, 2020, 2021)). We opted to focus on this slightly stylized scenario to maintain the clarity of our technical contribution. That said, as long as one is happy with a bound that scales with $C_{\varphi,1}$, a simple and elegant version of our algorithm can provide such bounds without prior knowledge of $\mathbf{\Lambda}$. Whether or not it is possible to unify the advantages of the two versions of our algorithm is an exciting question for future research.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 950180). M. Papini was supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

References

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

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.*, 9: 815–857, 2008.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9659–9668. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/uehara20a.html`.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. In *NeurIPS*, 2020.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Trans. Inf. Theory*, 68(12):8156–8196, 2022.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6683–6694. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/34f98c7c5d7063181da890ea8d25265a-Paper.pdf`.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3852–3878. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/cheng22b.html`.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2730–2775. PMLR, 02–05 Jul 2022. URL `https://proceedings.mlr.press/v178/zhan22a.html`.

Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, and Jiantao Jiao. Optimal conservative offline RL with general function approximation via augmented lagrangian. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/pdf?id=ZsvWb6mJnMv`.

Hanlin Zhu, Paria Rashidinejad, and Jiantao Jiao. Importance weighted actor-critic for optimal conservative offline reinforcement learning, 2023.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Alan S Manne. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960a.

Richard Bellman. Dynamic programming. Technical report, RAND CORP SANTA MONICA CA, 1956.

Prashant G. Mehta and Sean P. Meyn. Q-learning and pontryagin's minimum principle. In *CDC*, pages 3598–3605. IEEE, 2009.

Joan Bas-Serrano, Sebastian Curi, Andreas Krause, and Gergely Neu. Logistic q-learning. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3610–3618. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/bas-serrano21a.html`.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 6995–7004. PMLR, 2019.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 2020.

Mengdi Wang and Yichen Chen. An online primal-dual method for discounted markov decision processes. In *CDC*, pages 4516–4521. IEEE, 2016.

Yichen Chen, Lihong Li, and Mengdi Wang. Scalable bilinear learning using state and action features. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 833–842. PMLR, 2018.

Joan Bas-Serrano and Gergely Neu. Faster saddle-point optimization for solving large-scale markov decision processes. In *L4DC*, volume 120 of *Proceedings of Machine Learning Research*, pages 413–423. PMLR, 2020.

Gergely Neu and Nneka Okolo. Efficient global planning in large MDPs via stochastic primal-dual optimization. In *ALT*, volume 201 of *Proceedings of Machine Learning Research*, pages 1101–1123. PMLR, 2023.

Ofir Nachum and Bo Dai. Reinforcement learning via fenchel-rockafellar duality. 2020.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1994. ISBN 0471619779.

Richard Bellman. Dynamic programming. *Science*, 153 (3731):34–37, 1966.

Alan S. Manne. Linear programming and sequential decisions. *Manage. Sci.*, 6(3):259–267, apr 1960b. ISSN 0025-1909. doi: 10.1287/mnsc.6.3.259. URL `https://doi.org/10.1287/mnsc.6.3.259`.

Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982. ISBN 978-0-12-093480-5.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *ICLR*. OpenReview.net, 2022.

S.P. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1996.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *J. Mach. Learn. Res.*, 6:503–556, 2005.

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

Andrea Zanette, Martin J. Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. In *NeurIPS*, pages 13626–13640, 2021.

Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. In *AISTATS*, volume 151 of *Proceedings of Machine Learning Research*, pages 5757–5773. PMLR, 2022.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *NeurIPS*, pages 442–450. Curran Associates, Inc., 2010.

Lihong Li, Rémi Munos, and Csaba Szepesvári. On minimax optimal offline policy evaluation. *CoRR*, abs/1409.3653, 2014.

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *COLT*, volume 134 of *Proceedings of Machine Learning Research*, pages 3242–3245. PMLR, 2021.

Chenjun Xiao, Yifan Wu, Jincheng Mei, Bo Dai, Tor Lattimore, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. On the optimality of batch policy optimization algorithms. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 11362–11371. PMLR, 2021.

GM Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

Gergely Neu and Julia Olkhovskaya. Efficient and robust algorithms for adversarial linear contextual

bandits. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 3049–3068. PMLR, 2020.

Gergely Neu and Julia Olkhovskaya. Online learning in MDPs with linear function approximation and bandit feedback. In *NeurIPS*, pages 10407–10417, 2021.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2003.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Germano Gabbianelli, Gergely Neu, and Matteo Papini. Online learning with off-policy feedback. In Shipra Agrawal and Francesco Orabona, editors, *ALT*, volume 201 of *Proceedings of Machine Learning Research*, pages 620–641. PMLR, 20 Feb–23 Feb 2023. URL `https://proceedings.mlr.press/v201/gabbianelli23a.html`.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. No

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Not Applicable
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not Applicable

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Supplementary Material

A COMPLETE STATEMENT OF THEOREM 3.2

Theorem A.1. *Consider a linear MDP (Definition 2.1) such that $\boldsymbol{\theta}^\pi \in \mathbb{B}(D_{\boldsymbol{\theta}})$ for all $\pi \in \Pi$. Further, suppose that $C_{\boldsymbol{\varphi},c}(\pi^*;\pi_B) \le D^2_{\boldsymbol{\beta}}$. Then, for any comparator policy $\pi^* \in \Pi$, the policy output by Algorithm 1 satisfies:*

$$\mathbb{E}\left[\langle \boldsymbol{\mu}^{\pi^*} - \boldsymbol{\mu}^{\boldsymbol{\pi}_{out}}, \boldsymbol{r}\rangle\right] \le \frac{2D^2_{\boldsymbol{\beta}}}{\zeta T} + \frac{\log|\mathcal{A}|}{\alpha T} + \frac{2D^2_{\boldsymbol{\theta}}}{\eta K} + \frac{\zeta G^2_{\boldsymbol{\beta},c}}{2} + \frac{\alpha D^2_{\boldsymbol{\theta}} D^2_{\boldsymbol{\varphi}}}{2} + \frac{\eta G^2_{\boldsymbol{\theta},c}}{2},$$

where:

$$G^2_{\boldsymbol{\theta},c} = 3D^2_{\boldsymbol{\varphi}}\left((1-\gamma)^2 + (1+\gamma^2)D^2_{\boldsymbol{\beta}}\left\|\boldsymbol{\Lambda}\right\|_2^{2c-1}\right), \tag{18}$$

$$G^2_{\boldsymbol{\beta},c} = 3(1+(1+\gamma^2)D^2_{\boldsymbol{\varphi}}D^2_{\boldsymbol{\theta}})D^{2(2c-1)}_{\boldsymbol{\varphi}}. \tag{19}$$

In particular, using learning rates $\eta = \frac{2D_{\boldsymbol{\theta}}}{G_{\boldsymbol{\theta},c}\sqrt{K}}$, $\zeta = \frac{2D_{\boldsymbol{\beta}}}{G_{\boldsymbol{\beta},c}\sqrt{T}}$, and $\alpha = \frac{\sqrt{2\log|\mathcal{A}|}}{D_{\boldsymbol{\varphi}}D_{\boldsymbol{\theta}}\sqrt{T}}$, and setting $K = T\cdot\frac{2D_{\boldsymbol{\beta}^2}G^2_{\boldsymbol{\beta},c}+D^2_{\boldsymbol{\theta}}D^2_{\boldsymbol{\varphi}}\log|\mathcal{A}|}{2D^2_{\boldsymbol{\theta}}G^2_{\boldsymbol{\theta},c}}$, we achieve $\mathbb{E}\left[\langle \boldsymbol{\mu}^{\pi^*} - \boldsymbol{\mu}^{\boldsymbol{\pi}_{out}}, \boldsymbol{r}\rangle\right] \le \epsilon$ with a number of samples n_ϵ that is

$$O\left(\epsilon^{-4}D^4_{\boldsymbol{\theta}}D^4_{\boldsymbol{\varphi}}D^4_{\boldsymbol{\beta}}\operatorname{Tr}(\boldsymbol{\Lambda}^{2c-1})\left\|\boldsymbol{\Lambda}\right\|_2^{2c-1}\log|\mathcal{A}|\right).$$

By remark A.2 below, we have that n_ϵ is simply of order $O\left(\varepsilon^{-4}D^4_{\boldsymbol{\theta}}D^{8c}_{\boldsymbol{\varphi}}D^4_{\boldsymbol{\beta}}d^{2-2c}\log|\mathcal{A}|\right)$

Remark A.2. When $c = 1/2$, the factor $\operatorname{Tr}(\boldsymbol{\Lambda}^{2c-1})$ is just d, the feature dimension, and $\left\|\boldsymbol{\Lambda}\right\|_2^{2c-1} = 1$. When $c = 1$ and $\boldsymbol{\Lambda}$ is unknown, both $\left\|\boldsymbol{\Lambda}\right\|_2$ and $\operatorname{Tr}(\boldsymbol{\Lambda})$ should be replaced by their upper bound $D^2_{\boldsymbol{\varphi}}$. Then, for $c \in \{1/2, 1\}$, we have that $\operatorname{Tr}(\boldsymbol{\Lambda}^{2c-1})\left\|\boldsymbol{\Lambda}\right\|_2^{2c-1} \le D^{8c-4}_{\boldsymbol{\varphi}}d^{2-2c}$.

B MISSING PROOFS FOR THE DISCOUNTED SETTING

B.1 Proof of Lemma 4.1

Using the choice of comparators described in the lemma, we have

$$\begin{aligned}\nu_{\boldsymbol{\beta}^*}(s) &= (1-\gamma)\nu_0(s) + \gamma\langle \boldsymbol{\psi}(s), \boldsymbol{\Lambda}^c\boldsymbol{\Lambda}^{-c}\boldsymbol{\Phi}^\top\boldsymbol{\mu}^{\pi^*}\rangle \\ &= (1-\gamma)\nu_0(s) + \sum_{s',a'} p(s|s',a')\mu^{\pi^*}(s',a') = \nu^{\pi^*}(s),\end{aligned}$$

hence $\mu_{\boldsymbol{\beta}^*,\pi^*} = \boldsymbol{\mu}^{\pi^*}$. From Equation (14) it is easy to see that

$$\begin{aligned}f(\boldsymbol{\beta}^*, \pi^*; \boldsymbol{\theta}_t) &= \langle \boldsymbol{\Lambda}^{-c}\boldsymbol{\Phi}^\top\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^c\boldsymbol{\omega}\rangle + \langle \boldsymbol{\theta}_t, \boldsymbol{\Phi}^\top\boldsymbol{\mu}^* - \boldsymbol{\Lambda}^c\boldsymbol{\Lambda}^{-c}\boldsymbol{\Phi}^\top\boldsymbol{\mu}^*\rangle \\ &= \langle \boldsymbol{\mu}^{\pi^*}, \boldsymbol{\Phi}\boldsymbol{\omega}\rangle = \langle \boldsymbol{\mu}^*, \boldsymbol{r}\rangle.\end{aligned}$$

Moreover, we also have

$$\begin{aligned}v_{\boldsymbol{\theta}_t^*,\pi_t}(s) &= \sum_a \pi_t(a|s)\langle \boldsymbol{\theta}^{\pi_t}, \boldsymbol{\varphi}(x,a)\rangle \\ &= \sum_a \pi_t(a|s) q^{\pi_t}(s,a) = v^{\pi_t}(s,a).\end{aligned}$$

Then, from Equation (15) we obtain

$$\begin{aligned}&f(\boldsymbol{\beta}_t, \pi_t, \boldsymbol{\theta}_t^*) \\ &= (1-\gamma)\langle \boldsymbol{\nu}_0, v^{\pi_t}\rangle + \langle \boldsymbol{\beta}_t, \boldsymbol{\Lambda}^c(\boldsymbol{\omega} + \gamma\boldsymbol{\Psi}\boldsymbol{v}^{\pi_t} - \boldsymbol{\theta}^{\pi_t})\rangle \\ &= (1-\gamma)\langle \boldsymbol{\nu}_0, v^{\pi_t}\rangle + \langle \boldsymbol{\beta}_t, \boldsymbol{\Lambda}^{c-1}\mathbb{E}_{X,A\sim\mu_B}\left[\boldsymbol{\varphi}(X,A)\boldsymbol{\varphi}(X,A)^\intercal(\boldsymbol{\omega} + \gamma\boldsymbol{\Psi}\boldsymbol{v}^{\pi_t} - \boldsymbol{\theta}^{\pi_t})\right]\rangle \\ &= (1-\gamma)\langle \boldsymbol{\nu}_0, v^{\pi_t}\rangle + \langle \boldsymbol{\beta}_t, \boldsymbol{\Lambda}^{c-1}\mathbb{E}_{X,A\sim\mu_B}\left[[r(X,A) + \gamma\left\langle p(\cdot|X,A), \boldsymbol{v}^{\pi_t}\right\rangle - \boldsymbol{q}^{\pi_t}(X,A)]\boldsymbol{\varphi}(X,A)\right]\rangle \\ &= (1-\gamma)\langle \boldsymbol{\nu}_0, v^{\pi_t}\rangle = \langle \mu^{\pi_t}, \boldsymbol{r}\rangle,\end{aligned}$$

where the fourth equality uses that the value functions satisfy the Bellman equation $\boldsymbol{q}^\pi = \boldsymbol{r} + \gamma\boldsymbol{P}\boldsymbol{v}^\pi$ for any policy π. The proof is concluded by noticing that, since $\boldsymbol{\pi}_{\text{out}}$ is sampled uniformly from $\{\pi_t\}_{t=1}^T$, $\mathbb{E}\left[\langle \boldsymbol{\mu}^{\boldsymbol{\pi}_{\text{out}}}, \boldsymbol{r}\rangle\right] = \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\langle \boldsymbol{\mu}^{\pi_t}, \boldsymbol{r}\rangle\right]$. □

B.2 Proof of Lemma 4.2

We start by rewriting the terms appearing in the definition of $\mathcal{G}_T$:

$$\begin{aligned}f(\boldsymbol{\beta}^*, \pi^*; \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}_t, \pi_t; \boldsymbol{\theta}_t^*) &= f(\boldsymbol{\beta}^*, \pi^*; \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}^*, \pi_t; \boldsymbol{\theta}_t) \\ &\quad + f(\boldsymbol{\beta}^*, \pi_t; \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}_t, \pi_t; \boldsymbol{\theta}_t) \\ &\quad + f(\boldsymbol{\beta}_t, \pi_t; \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}_t, \pi_t; \boldsymbol{\theta}_t^*). \qquad (20)\end{aligned}$$

To rewrite this as the sum of the three regret terms, we first note that

$$f(\boldsymbol{\beta}, \pi; \boldsymbol{\theta}) = \langle \boldsymbol{\Lambda}^c\boldsymbol{\beta}, \boldsymbol{\omega} - \boldsymbol{\theta}_t\rangle + \langle \nu_{\boldsymbol{\beta}}, v_{\boldsymbol{\theta}_t,\pi}\rangle,$$

which allows us to write the first term of Equation (20) as

$$\begin{aligned}f(\boldsymbol{\beta}^*, \pi^*; \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}^*, \pi_t; \boldsymbol{\theta}_t) &= \langle \boldsymbol{\Lambda}^c(\boldsymbol{\beta}^* - \boldsymbol{\beta}^*), \boldsymbol{\omega} - \boldsymbol{\theta}_t\rangle + \langle \nu_{\boldsymbol{\beta}^*}, v_{\boldsymbol{\theta}_t,\pi^*} - v_{\boldsymbol{\theta}_t,\pi_t}\rangle \\ &= \langle \nu_{\boldsymbol{\beta}^*}, \sum_a (\pi^*(a|\cdot) - \pi_t(a|\cdot))\langle \boldsymbol{\theta}_t, \boldsymbol{\varphi}(\cdot,a)\rangle\rangle,\end{aligned}$$

and we have already established in the proof of Lemma C.3 that $\boldsymbol{\nu}_{\boldsymbol{\beta}^*}$ is equal to $\boldsymbol{\nu}^{\pi^*}$ for our choice of comparator. Similarly, we use Equation (15) to rewrite the second term of Equation (20) as

$$\begin{aligned}f(\boldsymbol{\beta}^*, \pi_t; \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}_t, \pi_t; \boldsymbol{\theta}_t) &= (1-\gamma)\langle \boldsymbol{\nu}_0, v_{\boldsymbol{\theta}_t,\pi_t} - v_{\boldsymbol{\theta}_t,\pi_t}\rangle + \langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_t, \boldsymbol{\Lambda}^c(\boldsymbol{\omega} + \gamma\boldsymbol{\Psi}v_{\boldsymbol{\theta}_t,\pi_t} - \boldsymbol{\theta}_t)\rangle \\ &= \langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_t, g_{\boldsymbol{\beta},t}\rangle.\end{aligned}$$

Finally, we use Equation (14) to rewrite the third term of Equation (20) as

$$\begin{aligned}f(\boldsymbol{\beta}_t, \pi_t; \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}_t, \pi_t; \boldsymbol{\theta}_t^*) &= \langle \boldsymbol{\beta}_t - \boldsymbol{\beta}_t, \boldsymbol{\Lambda}^c\boldsymbol{\omega}\rangle + \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*, \boldsymbol{\Phi}^\top\boldsymbol{\mu}_{\boldsymbol{\beta}_t,\pi_t} - \boldsymbol{\Lambda}^c\boldsymbol{\beta}_t\rangle \\ &= \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*, g_{\boldsymbol{\theta},t}\rangle.\end{aligned}$$

B.3 Regret bounds for stochastic gradient descent / ascent

Lemma B.1. *For any dynamic comparator $\boldsymbol{\theta}_{1:T} \in D_{\boldsymbol{\theta}}$, the iterates $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_T$ of Algorithm 1 satisfy the following regret bound:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*, g_{\boldsymbol{\theta},t} \rangle\right] \leq \frac{2TD_{\boldsymbol{\theta}}^2}{\eta K} + \frac{3\eta T D_{\boldsymbol{\varphi}}^2 \left((1-\gamma)^2 + (1+\gamma^2) D_{\beta}^2 \left\|\boldsymbol{\Lambda}\right\|_2^{2c-1}\right)}{2}.$$

Proof. First, we use the definition of $\boldsymbol{\theta}_t$ as the average of the inner-loop iterates from Algorithm 1, together with linearity of expectation and bilinearity of the inner product.

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*, g_{\boldsymbol{\theta},t} \rangle\right] = \sum_{t=1}^{T} \frac{1}{K} \underbrace{\mathbb{E}\left[\sum_{k=1}^{K} \langle \boldsymbol{\theta}_{t,k} - \boldsymbol{\theta}_t^*, g_{\boldsymbol{\theta},t} \rangle\right]}_{\mathfrak{R}_t}. \tag{21}$$

We then appeal to standard stochastic gradient descent analysis to bound each term $\mathfrak{R}_t$ separately.

We have already proven in Section 4 that the gradient estimator for $\boldsymbol{\theta}$ is unbiased, that is, $\mathbb{E}_{t,k}\left[\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,k}\right] = \boldsymbol{g}_{\theta,t}$. It is also useful to recall here that $\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,k}$ does *not* depend on $\boldsymbol{\theta}_{t,k}$. Next, we show that its second moment is bounded. From Equation (11), plugging in the definition of $\mu_{t,k}$ from Equation (9) and using the abbreviations $\boldsymbol{\varphi}_{t,k}^0 = \sum_a \pi_t(a|x_{t,k}^0)\boldsymbol{\varphi}(x_{t,k}^0, a)$, $\boldsymbol{\varphi}_t = \boldsymbol{\varphi}(x_{t,k}, a_{t,k})$, and $\boldsymbol{\varphi}'_{t,k} = \sum_a \pi_t(a|x'_{t,k})\boldsymbol{\varphi}(x'_{t,k}, a)$, we have:

$$\begin{aligned}
&\mathbb{E}_{t,k}\left[\left\|\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,i}\right\|^2\right] \\
&\quad= \mathbb{E}_{t,k}\left[\left\|(1-\gamma)\boldsymbol{\varphi}_{t,k}^0 + \gamma\boldsymbol{\varphi}'_{t,k}\langle \boldsymbol{\varphi}_{tk}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\rangle - \boldsymbol{\varphi}_{t,k}\langle \boldsymbol{\varphi}_{tk}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\rangle\right\|^2\right] \\
&\quad\leq 3(1-\gamma)^2\mathcal{D}_{\boldsymbol{\varphi}}^2 + 3\gamma^2\mathbb{E}_{t,k}\left[\left\|\boldsymbol{\varphi}'_{t,k}\langle \boldsymbol{\varphi}_{tk}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\rangle\right\|^2\right] + 3\mathbb{E}_{t,k}\left[\left\|\boldsymbol{\varphi}_{t,k}\langle \boldsymbol{\varphi}_{tk}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\rangle\right\|^2\right] \\
&\quad\leq 3(1-\gamma)^2\mathcal{D}_{\boldsymbol{\varphi}}^2 + 3(1+\gamma^2)D_{\boldsymbol{\varphi}}^2\mathbb{E}_{t,k}\left[\langle \boldsymbol{\varphi}_{tk}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\rangle^2\right] \\
&\quad= 3(1-\gamma)^2\mathcal{D}_{\boldsymbol{\varphi}}^2 + 3(1+\gamma^2)D_{\boldsymbol{\varphi}}^2\boldsymbol{\beta}_t^\top\boldsymbol{\Lambda}^{c-1}\mathbb{E}_{t,k}\left[\boldsymbol{\varphi}_{tk}\boldsymbol{\varphi}_{tk}^\top\right]\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t \\
&\quad= 3(1-\gamma)^2\mathcal{D}_{\boldsymbol{\varphi}}^2 + 3(1+\gamma^2)D_{\boldsymbol{\varphi}}^2\left\|\boldsymbol{\beta}_t\right\|_{\boldsymbol{\Lambda}^{2c-1}}^2.
\end{aligned}$$

We can then apply Lemma D.1 with the latter expression as G^2, $\mathbb{B}(D_{\boldsymbol{\theta}})$ as the domain, and η as the learning rate, obtaining:

$$\begin{aligned}
\mathbb{E}_t\left[\sum_{k=1}^{K} \langle \boldsymbol{\theta}_{t,k} - \boldsymbol{\theta}_t^*, g_{\boldsymbol{\theta},t} \rangle\right] &\leq \frac{\left\|\boldsymbol{\theta}_{t,1} - \boldsymbol{\theta}_t^*\right\|_2^2}{2\eta} + \frac{3\eta K D_{\boldsymbol{\varphi}}^2 \left((1-\gamma)^2 + (1+\gamma^2)\left\|\boldsymbol{\beta}_t\right\|_{\boldsymbol{\Lambda}^{2c-1}}^2\right)}{2} \\
&\leq \frac{2D_{\theta}^2}{\eta} + \frac{3\eta K D_{\boldsymbol{\varphi}}^2 \left((1-\gamma)^2 + (1+\gamma^2)\left\|\boldsymbol{\beta}_t\right\|_{\boldsymbol{\Lambda}^{2c-1}}^2\right)}{2}.
\end{aligned}$$

Plugging this into Equation (21) and bounding $\left\|\boldsymbol{\beta}_t\right\|_{\boldsymbol{\Lambda}^{2c-1}}^2 \leq D_{\boldsymbol{\beta}}^2\left\|\boldsymbol{\Lambda}\right\|_2^{2c-1}$, we obtain the final result. □

Lemma B.2. *For any comparator $\boldsymbol{\beta} \in D_{\boldsymbol{\beta}}$, the iterates $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_T$ of Algorithm 1 satisfy the following regret bound:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_t, g_{\boldsymbol{\beta},t} \rangle\right] \leq \frac{2D_{\boldsymbol{\beta}}^2}{\zeta} + \frac{3\zeta T(1 + (1+\gamma^2)D_{\boldsymbol{\varphi}}^2 D_{\boldsymbol{\theta}}^2)\operatorname{Tr}(\boldsymbol{\Lambda}^{2c-1})}{2}.$$

Proof. We again employ stochastic gradient descent analysis. We first prove that the gradient estimator for $\boldsymbol{\beta}$ is

unbiased. Recalling the definition of $\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t}$ from Equation (10),

$$\begin{aligned}
\mathbb{E}\left[\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t}|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right] &= \mathbb{E}\left[\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_t\left(R_t+\gamma v_t(X_t')-\langle\boldsymbol{\varphi}_t,\boldsymbol{\theta}_t\rangle\right)|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right] \\
&= \boldsymbol{\Lambda}^{c-1}\left(\mathbb{E}_t\left[\boldsymbol{\varphi}_t\boldsymbol{\varphi}_t^\top\right]\boldsymbol{\omega}+\gamma\mathbb{E}_t\left[\boldsymbol{\varphi}_t v_t(X_t')\right]-\mathbb{E}_t\left[\boldsymbol{\varphi}_t\boldsymbol{\varphi}_t^\top\right]\boldsymbol{\theta}_t\right) \\
&= \boldsymbol{\Lambda}^{c-1}\left(\boldsymbol{\Lambda}\boldsymbol{\omega}+\gamma\mathbb{E}_t\left[\boldsymbol{\varphi}_t v_t(X_t')\right]-\boldsymbol{\Lambda}\boldsymbol{\theta}_t\right) \\
&= \boldsymbol{\Lambda}^{c-1}\left(\boldsymbol{\Lambda}\boldsymbol{\omega}+\gamma\mathbb{E}_t\left[\boldsymbol{\varphi}_t\boldsymbol{P}(\cdot|X_t,A_t)\boldsymbol{v}_t\right]-\boldsymbol{\Lambda}\boldsymbol{\theta}_t\right) \\
&= \boldsymbol{\Lambda}^{c-1}\left(\boldsymbol{\Lambda}\boldsymbol{\omega}+\gamma\mathbb{E}_t\left[\boldsymbol{\varphi}_t\boldsymbol{\varphi}_t^\top\right]\boldsymbol{\Psi}\boldsymbol{v}_t-\boldsymbol{\Lambda}\boldsymbol{\theta}_t\right) \\
&= \boldsymbol{\Lambda}^{c}(\boldsymbol{\omega}+\gamma\boldsymbol{\Psi}v_{\boldsymbol{\theta}_t,\pi_t}-\boldsymbol{\theta}_t) = \boldsymbol{g}_{\beta,t},
\end{aligned}$$

recalling that $\boldsymbol{v}_t = \boldsymbol{v}_{\boldsymbol{\theta}_t,\pi_t}$. Next, we bound its second moment. We use the fact that $r \in [0,1]$ and $\|\boldsymbol{v}_t\|_\infty \le \|\boldsymbol{\Phi}\boldsymbol{\theta}_t\|_\infty \le D_{\boldsymbol{\varphi}}D_{\boldsymbol{\theta}}$ to show that

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t}\|_2^2|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right] &= \mathbb{E}\left[\left\|\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_t[R_t+\gamma v_t(X_t')-\langle\boldsymbol{\theta}_t,\boldsymbol{\varphi}_t\rangle]\right\|_2^2|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right] \\
&\le 3(1+(1+\gamma^2)D_{\boldsymbol{\varphi}}^2D_{\boldsymbol{\theta}}^2)\mathbb{E}_t\left[\boldsymbol{\varphi}_t^\intercal\boldsymbol{\Lambda}^{2(c-1)}\boldsymbol{\varphi}_t\right] \\
&= 3(1+(1+\gamma^2)D_{\boldsymbol{\varphi}}^2D_{\boldsymbol{\theta}}^2)\mathbb{E}_t\left[\mathrm{Tr}(\boldsymbol{\Lambda}^{2(c-1)}\boldsymbol{\varphi}_t\boldsymbol{\varphi}_t^\intercal)\right] \\
&= 3(1+(1+\gamma^2)D_{\boldsymbol{\varphi}}^2D_{\boldsymbol{\theta}}^2)\,\mathrm{Tr}(\boldsymbol{\Lambda}^{2c-1}).
\end{aligned}$$

Thus, we can apply Lemma D.1 with the latter expression as G^2, $\mathbb{B}(D_{\boldsymbol{\beta}})$ as the domain, and ζ as the learning rate. □

Lemma B.3. *The sequence of policies $\pi_1,\dots,\pi_T$ of Algorithm 1 satisfies the following regret bound:*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x\in\mathcal{X}}\nu^{\pi^*}(x)\sum_a(\pi^*(a|x)-\pi_t(a|x))\langle\boldsymbol{\theta}_t,\boldsymbol{\varphi}(x,a)\rangle\right] \le \frac{\log|\mathcal{A}|}{\alpha}+\frac{\alpha T D_{\boldsymbol{\varphi}}^2D_{\boldsymbol{\theta}}^2}{2}.$$

Proof. We just apply mirror descent analysis, invoking Lemma D.2 with $q_t = \Phi\boldsymbol{\theta}_t$, noting that $\|q_t\|_\infty \le D_{\boldsymbol{\varphi}}D_{\boldsymbol{\theta}}$. The proof is concluded by trivially bounding the relative entropy as $\mathcal{H}(\pi^*\|\pi_1) = \mathbb{E}_{x\sim\nu^{\pi^*}}\left[\mathcal{D}(\pi(\cdot|x)\|\pi_1(\cdot|x))\right] \le \log|\mathcal{A}|$. □

C ANALYSIS FOR THE AVERAGE-REWARD MDP SETTING

This section describes the adaptation of our contributions in the main body of the paper to average-reward MDPs (AMDPs). In the offline reinforcement learning setting that we consider, we assume access to a sequence of data points (X_t, A_t, R_t, X_t') in round t generated by a behaviour policy π_B whose occupancy measure is denoted as $\boldsymbol{\mu}_B$. Specifically, we will now draw i.i.d. samples from the *undiscounted* occupancy measure as $X_t, A_t \sim \boldsymbol{\mu}_B$, sample $X_t' \sim p(\cdot|X_t, A_t)$, and compute immediate rewards as $R_t = r(X_t, A_t)$. For simplicity, we use the shorthand notation $\boldsymbol{\varphi}_t = \varphi(X_t, A_t)$ to denote the feature vector drawn in round t, and define the matrix $\boldsymbol{\Lambda} = \mathbb{E}\left[\varphi(X_t, A_t)\varphi(X_t, A_t)^\top\right]$.

Before describing our contributions, some definitions are in order. An important central concept in the theory of AMDPs is that of the *relative value functions* of policy π defined as

$$v^\pi(x) = \lim_{T\to\infty} \mathbb{E}_\pi\left[\sum_{t=0}^{T} r(X_t, A_t) - \rho^\pi \middle| X_0 = x\right],$$
$$q^\pi(x, a) = \lim_{T\to\infty} \mathbb{E}_\pi\left[\sum_{t=0}^{T} r(X_t, A_t) - \rho^\pi \middle| X_0 = x, A_0 = a\right],$$

where we recalled the notation ρ^π denoting the average reward of policy π from the main text. These functions are sometimes also called the *bias functions*, and their intuitive role is to measure the total amount of reward gathered by policy π before it hits its stationary distribution. For simplicity, we will refer to these functions as value functions and action-value functions below.

By their recursive nature, these value functions are also characterized by the corresponding Bellman equations recalled below for completeness

$$\boldsymbol{q}^\pi = \boldsymbol{r} - \rho^\pi \mathbf{1} + \boldsymbol{P}\boldsymbol{v}^\pi,$$

where $\boldsymbol{v}^\pi$ is related to the action-value function as $v^\pi(x) = \sum_a \pi(a|x) q^\pi(x, a)$. We note that the Bellman equations only characterize the value functions up to a constant offset. That is, for any policy π, and constant $c \in \mathbb{R}$, $\boldsymbol{v}^\pi + c\mathbf{1}$ and $\boldsymbol{q}^\pi + c\mathbf{1}$ also satisfy the Bellman equations. A key quantity to measure the size of the value functions is the *span seminorm* defined for $\boldsymbol{q} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ as $\|\boldsymbol{q}\|_{\mathrm{sp}} = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} q(x, a) - \inf_{(x,a)\in\mathcal{X}\times\mathcal{A}} q(x, a)$. Using this notation, the condition of Assumption 5.1 can be simply stated as requiring $\|\boldsymbol{q}^\pi\|_{\mathrm{sp}} \le D_q$ for all π.

Now, let π^* denote an optimal policy with maximum average reward and introduce the shorthand notations $\rho^* = \rho^{\pi^*}, \boldsymbol{\mu}^* = \boldsymbol{\mu}^{\pi^*}, \boldsymbol{\nu}^* = \boldsymbol{\nu}^{\pi^*}, \boldsymbol{v}^* = \boldsymbol{v}^{\pi^*}$ and $\boldsymbol{q}^* = \boldsymbol{q}^{\pi^*}$. Under mild assumptions on the MDP that we will clarify shortly, the following Bellman optimality equations are known to characterize bias vectors corresponding to the optimal policy

$$\boldsymbol{q}^* = \boldsymbol{r} - \rho^* \mathbf{1} + \boldsymbol{P}\boldsymbol{v}^*,$$

where $\boldsymbol{v}^*$ satisfies $v^*(x) = \max_a q^*(x, a)$. Once again, shifting the solutions by a constant preserves the optimality conditions. It is easy to see that such constant offsets do not influence greedy or softmax policies extracted from the action value functions. Importantly, by a calculation analogous to Equation (3), the action-value functions are exactly realizable under the linear MDP condition (see Definition 2.1) and Assumption 5.2.

Besides the Bellman optimality equations stated above, optimal policies can be equivalently characterized via the following linear program:

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{\mu}, \boldsymbol{r} \rangle \\ \text{subject to} \quad & \boldsymbol{E}^\top \boldsymbol{\mu} = \boldsymbol{P}^\top \boldsymbol{\mu} \\ & \langle \boldsymbol{\mu}, \mathbf{1} \rangle = 1 \\ & \boldsymbol{\mu} \ge 0. \end{aligned} \tag{22}$$

This can be seen as the generalization of the LP stated for discounted MDPs in the main text, with the added complication that we need to make sure that the occupancy measures are normalized[3] to 1. By following the same steps as in the main text to relax the constraints and reparametrize the LP, one can show that solutions of the

[3]This is necessary because of the absence of ν_0 in the LP, which would otherwise fix the scale of the solutions.

LP under the linear MDP assumption can be constructed by finding the saddle point of the following Lagrangian:

$$\begin{aligned}\mathfrak{L}(\boldsymbol{\lambda}, \boldsymbol{\mu}; \rho, \boldsymbol{v}, \boldsymbol{\theta}) &= \rho + \langle \boldsymbol{\lambda}\,, \boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v} - \boldsymbol{\theta} - \rho\boldsymbol{\varrho}\rangle + \langle \boldsymbol{u}\,, \boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{E}\boldsymbol{v}\rangle \\ &= \rho[1 - \langle \boldsymbol{\lambda}, \boldsymbol{\varrho}\rangle] + \langle \boldsymbol{\theta}, \boldsymbol{\Phi}^\intercal\boldsymbol{\mu} - \boldsymbol{\lambda}\rangle + \langle \boldsymbol{v}, \boldsymbol{\Psi}^\intercal\boldsymbol{\lambda} - \boldsymbol{E}^\intercal\boldsymbol{\mu}\rangle\,.\end{aligned}$$

As before, the optimal value functions $\boldsymbol{q}^*$ and $\boldsymbol{v}^*$ are optimal primal variables for the saddle-point problem, as are all of their constant shifts. Thus, the existence of a solution with small span seminorm implies the existence of a solution with small supremum norm.

Finally, applying the same reparametrization $\boldsymbol{\beta} = \boldsymbol{\Lambda}^{-c}\boldsymbol{\lambda}$ as in the discounted setting, we arrive to the following Lagrangian that forms the basis of our algorithm:

$$\mathfrak{L}(\boldsymbol{\beta}, \boldsymbol{\mu}; \rho, \boldsymbol{v}, \boldsymbol{\theta}) = \rho + \langle \boldsymbol{\beta}\,, \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v} - \boldsymbol{\theta} - \rho\boldsymbol{\varrho}]\rangle + \langle \boldsymbol{\mu}\,, \boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{E}\boldsymbol{v}\rangle.$$

We will aim to find the saddle point of this function via primal-dual methods. As we have some prior knowledge of the optimal solutions, we will restrict the search space of each optimization variable to nicely chosen compact sets. For the $\boldsymbol{\beta}$ iterates, we consider the Euclidean ball domain $\mathbb{B}(D_{\boldsymbol{\beta}}) = \{\boldsymbol{\beta} \in \mathbb{R}^d \mid \|\boldsymbol{\beta}\|_2 \le D_{\boldsymbol{\beta}}\}$ with the bound $D_{\boldsymbol{\beta}} > \|\boldsymbol{\Phi}^\intercal\boldsymbol{\mu}^*\|_{\boldsymbol{\Lambda}^{-2c}}$. Since the average reward of any policy is bounded in $[0, 1]$, we naturally restrict the ρ iterates to this domain. Finally, keeping in mind that Assumption 5.1 guarantees that $\|\boldsymbol{q}^\pi\|_{\mathrm{sp}} \le D_q$, we will also constrain the $\boldsymbol{\theta}$ iterates to an appropriate domain: $\mathbb{B}(D_{\boldsymbol{\theta}}) = \{\boldsymbol{\theta} \in \mathbb{R}^d \mid \|\boldsymbol{\theta}\|_2 \le D_{\boldsymbol{\theta}}\}$. We will assume that this domain is large enough to represent all action-value functions, which implies that $D_{\boldsymbol{\theta}}$ should scale at least linearly with D_q. Indeed, we will suppose that the features are bounded as $\|\boldsymbol{\varphi}(x, a)\|_2 \le D_{\boldsymbol{\varphi}}$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ so that our optimization algorithm only admits parametric $\boldsymbol{q}$ functions satisfying $\|\boldsymbol{q}\|_\infty \le D_{\boldsymbol{\varphi}} D_{\boldsymbol{\theta}}$. Obviously, $D_{\boldsymbol{\theta}}$ needs to be set large enough to ensure that it is possible at all to represent $\boldsymbol{q}$-functions with span D_q.

Thus, we aim to solve the following constrained optimization problem:

$$\min_{\rho\in[0,1], \boldsymbol{v}\in\mathbb{R}^{\mathcal{X}}, \boldsymbol{\theta}\in\mathbb{B}(D_{\boldsymbol{\theta}})} \max_{\boldsymbol{\beta}\in\mathbb{B}(D_{\boldsymbol{\beta}}), \boldsymbol{\mu}\in\mathbb{R}_+^{\mathcal{X}\times\mathcal{A}}} \mathfrak{L}(\boldsymbol{\beta}, \boldsymbol{\mu}; \rho, \boldsymbol{v}, \boldsymbol{\theta}).$$

As done in the main text, we eliminate the high-dimensional variables $\boldsymbol{v}$ and $\boldsymbol{\mu}$ by committing to the choices $\boldsymbol{v} = \boldsymbol{v}_{\boldsymbol{\theta},\pi}$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_{\boldsymbol{\beta},\pi}$ defined as

$$\begin{aligned}v_{\boldsymbol{\theta},\pi}(x) &= \sum_a \pi(a|x)\,\langle \boldsymbol{\theta}, \boldsymbol{\varphi}(x, a)\rangle\,, \\ \mu_{\boldsymbol{\beta},\pi}(x, a) &= \pi(a|x)\langle \boldsymbol{\psi}(x), \boldsymbol{\Lambda}^c\boldsymbol{\beta}\rangle.\end{aligned}$$

This makes it possible to express the Lagrangian in terms of only $\boldsymbol{\beta}, \pi, \rho$ and $\boldsymbol{\theta}$:

$$\begin{aligned}f(\boldsymbol{\beta}, \pi; \rho, \boldsymbol{\theta}) &= \rho + \langle \boldsymbol{\beta}\,, \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\theta},\pi} - \boldsymbol{\theta} - \rho\boldsymbol{\varrho}]\rangle + \langle \boldsymbol{\mu}_{\boldsymbol{\beta},\pi}\,, \boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{E}\boldsymbol{v}_{\boldsymbol{\theta},\pi}\rangle \\ &= \rho + \langle \boldsymbol{\beta}\,, \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\theta},\pi} - \boldsymbol{\theta} - \rho\boldsymbol{\varrho}]\rangle\end{aligned}$$

The remaining low-dimensional variables $\boldsymbol{\beta}, \rho, \boldsymbol{\theta}$ are then updated using stochastic gradient descent/ascent. For this purpose it is useful to express the partial derivatives of the Lagrangian with respect to said variables:

$$\begin{aligned}\boldsymbol{g}_{\boldsymbol{\beta}} &= \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\theta},\pi} - \boldsymbol{\theta} - \rho\boldsymbol{\varrho}] \\ g_\rho &= 1 - \langle \boldsymbol{\beta}, \boldsymbol{\Lambda}^c\boldsymbol{\varrho}\rangle \\ \boldsymbol{g}_{\boldsymbol{\theta}} &= \boldsymbol{\Phi}^\intercal\boldsymbol{\mu}_{\boldsymbol{\beta},\pi} - \boldsymbol{\Lambda}^c\boldsymbol{\beta}\end{aligned}$$

C.1 Algorithm for average-reward MDPs

Our algorithm for the AMDP setting has the same double-loop structure as the one for the discounted setting. In particular, the algorithm performs a sequence of outer updates $t = 1, 2, \ldots, T$ on the policies π_t and the iterates $\boldsymbol{\beta}_t$, and then performs a sequence of updates $i = 1, 2, \ldots, K$ in the inner loop to evaluate the policies and produce $\boldsymbol{\theta}_t$, ρ_t and $\boldsymbol{v}_t$. Thanks to the reparametrization $\boldsymbol{\beta} = \boldsymbol{\Lambda}^{-c}\boldsymbol{\lambda}$, fixing $\pi_t = \mathrm{softmax}(\sum_{k=1}^{t-1} \boldsymbol{\Phi}\boldsymbol{\theta}_k)$, $\boldsymbol{v}_t(x) = \sum_{a\in\mathcal{A}} \pi_t(a|x)\,\langle \varphi(x, a), \boldsymbol{\theta}_t\rangle$ for $x \in \mathcal{X}$, and $\mu_t(x, a) = \pi_t(a|x)\,\langle \boldsymbol{\psi}(x), \boldsymbol{\Lambda}^c\boldsymbol{\beta}_t\rangle$ in round t we can obtain unbiased estimates of the gradients of f with respect to $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and ρ. For each primal update t, the algorithm uses a single sample transition (X_t, A_t, R_t, X_t') generated by the behavior policy π_B to compute an unbiased

Algorithm 2 Offline primal-dual method for Average-reward MDPs

Input: Learning rates ζ, α,ξ,η, initial iterates $\boldsymbol{\beta}_1 \in \mathbb{B}(D_{\boldsymbol{\beta}})$, $\rho_0 \in [0,1]$, $\boldsymbol{\theta}_0 \in \mathbb{B}(D_{\boldsymbol{\theta}})$, $\pi_1 \in \Pi$,

for $t = 1$ **to** T **do**
  // Stochastic gradient descent:
  Initialize: $\boldsymbol{\theta}_t^{(1)} = \boldsymbol{\theta}_{t-1}$;
  for $i = 1$ **to** K **do**
    Obtain sample $W_{t,i} = (X_{t,i}, A_{t,i}, R_{t,i}, X'_{t,i})$;
    Sample $A'_{t,i} \sim \pi_t(\cdot|X'_{t,i})$;

    Compute $\tilde{g}_{\rho,t,i} = 1 - \left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t \right\rangle$;
    $\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,i} = \boldsymbol{\varphi}'_{t,i} \left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t \right\rangle - \boldsymbol{\varphi}_{t,i} \left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t \right\rangle$;

    Update $\rho_t^{(i+1)} = \Pi_{[0,1]}(\rho_t^{(i)} - \xi\tilde{g}_{\rho,t,i})$;
    $\boldsymbol{\theta}_t^{(i+1)} = \Pi_{\mathbb{B}(D_{\boldsymbol{\theta}})}(\boldsymbol{\theta}_t^{(i)} - \eta\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,i})$.
  end for
  Compute $\rho_t = \dfrac{1}{K}\sum_{i=1}^K \rho_t^{(i)}$;
  $\boldsymbol{\theta}_t = \dfrac{1}{K}\sum_{i=1}^K \boldsymbol{\theta}_t^{(i)}$;

  // Stochastic gradient ascent:
  Obtain sample $W_t = (X_t, A_t, R_t, X'_t)$;
  Compute $v_t(X'_t) = \sum_a \pi_t(a|X'_t)\langle \boldsymbol{\varphi}(X'_t, a), \boldsymbol{\theta}_t\rangle$;
  Compute $\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t} = \boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_t[R_t + v_t(X'_t) - \langle \boldsymbol{\theta}_t, \boldsymbol{\varphi}_t\rangle - \rho_t]$;
  Update $\boldsymbol{\beta}_{t+1} = \Pi_{\mathbb{B}(D_{\boldsymbol{\beta}})}(\boldsymbol{\beta}_t + \zeta\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t})$;

  // Policy update:
  Compute $\pi_{t+1} = \sigma\left(\alpha\sum_{k=1}^t \boldsymbol{\Phi}\boldsymbol{\theta}_k\right)$.
end for
Return: π_J with $J \sim \mathcal{U}(T)$.

estimator of the first gradient $g_{\boldsymbol{\beta}}$ for that round as $\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t} = \boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_t[R_t + v_t(X'_t) - \langle \boldsymbol{\theta}_t, \boldsymbol{\varphi}_t\rangle - \rho_t]$. Then, in iteration $i = 1, \cdots, K$ of the inner loop within round t, we sample transitions $(X_{t,i}, A_{t,i}, R_{t,i}, X'_{t,i})$ to compute gradient estimators with respect to ρ and $\boldsymbol{\theta}$ as:

$$\begin{aligned}
\tilde{g}_{\rho,t,i} &= 1 - \left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t \right\rangle \\
\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,i} &= \boldsymbol{\varphi}'_{t,i} \left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t \right\rangle - \boldsymbol{\varphi}_{t,i} \left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t \right\rangle.
\end{aligned}$$

We have used the shorthand notation $\boldsymbol{\varphi}_{t,i} = \varphi(X_{t,i}, A_{t,i})$, $\boldsymbol{\varphi}'_{t,i} = \varphi(X'_{t,i}, A'_{t,i})$. The update steps are detailed in the pseudocode presented as Algorithm 2.

We now state the general form of our main result for this setting in Theorem C.1 below.

Theorem C.1. *Consider a linear MDP (Definition 2.1) such that $\boldsymbol{\theta}^\pi \in \mathbb{B}(D_{\boldsymbol{\theta}})$ for all $\pi \in \Pi$. Further, suppose that $C_{\varphi,c}(\pi^*; \pi_B) \le D_{\boldsymbol{\beta}}^2$. Then, for any comparator policy $\pi^* \in \Pi$, the policy output by Algorithm 2 satisfies:*

$$\mathbb{E}\left[\left\langle \boldsymbol{\mu}^{\pi^*} - \boldsymbol{\mu}^{\boldsymbol{\pi}_{out}}, \boldsymbol{r}\right\rangle\right] \le \frac{2D_{\boldsymbol{\beta}}^2}{\zeta T} + \frac{\log|\mathcal{A}|}{\alpha T} + \frac{1}{2\xi K} + \frac{2D_{\boldsymbol{\theta}}^2}{\eta K} + \frac{\zeta G_{\boldsymbol{\beta},c}^2}{2} + \frac{\alpha D_{\boldsymbol{\theta}}^2 D_{\boldsymbol{\varphi}}^2}{2} + \frac{\xi G_{\rho,c}^2}{2} + \frac{\eta G_{\boldsymbol{\theta},c}^2}{2},$$

where

$$G_{\boldsymbol{\beta},c}^2 = \mathrm{Tr}(\boldsymbol{\Lambda}^{2c-1})(1 + 2D_{\boldsymbol{\theta}}D_{\boldsymbol{\varphi}})^2, \tag{23}$$

$$G_{\rho,c}^2 = 2\left(1 + D_{\boldsymbol{\beta}}^2 \|\boldsymbol{\Lambda}\|_2^{2c-1}\right), \tag{24}$$

$$G_{\boldsymbol{\theta},c}^2 = 4D_{\boldsymbol{\varphi}}^2 D_{\boldsymbol{\beta}}^2 \|\boldsymbol{\Lambda}\|_2^{2c-1}. \tag{25}$$

In particular, using learning rates $\zeta = \frac{2D_{\boldsymbol{\beta}}}{G_{\boldsymbol{\beta},c}\sqrt{T}}$, $\alpha = \frac{\sqrt{2\log|\mathcal{A}|}}{D_{\boldsymbol{\theta}}D_{\boldsymbol{\varphi}}\sqrt{T}}$, $\xi = \frac{1}{G_{\rho,c}\sqrt{K}}$, *and* $\eta = \frac{2D_{\boldsymbol{\theta}}}{G_{\boldsymbol{\theta},c}\sqrt{K}}$, *and setting* $K = T \cdot \frac{4D_{\boldsymbol{\beta}^2}G^2_{\boldsymbol{\beta},c}+2D^2_{\boldsymbol{\theta}}D^2_{\boldsymbol{\varphi}}\log|\mathcal{A}|}{G^2_{\rho,c}+4D^2_{\boldsymbol{\theta}}G^2_{\boldsymbol{\theta},c}}$, *we achieve* $\mathbb{E}\left[\langle \boldsymbol{\mu}^{\pi^*} - \boldsymbol{\mu}^{\boldsymbol{\pi}_{out}}, \boldsymbol{r}\rangle\right] \le \epsilon$ *with a number of samples* n_ϵ *that is*

$$O\left(\epsilon^{-4} D^4_{\boldsymbol{\theta}} D^4_{\boldsymbol{\varphi}} D^4_{\boldsymbol{\beta}} \operatorname{Tr}(\boldsymbol{\Lambda}^{2c-1}) \|\boldsymbol{\Lambda}\|_2^{2(2c-1)} \log|\mathcal{A}|\right).$$

By remark A.2, we have that n_ϵ is of order $O\left(\varepsilon^{-4} D^4_{\boldsymbol{\theta}} D^{12c-2}_{\boldsymbol{\varphi}} D^4_{\boldsymbol{\beta}} d^{2-2c} \log|\mathcal{A}|\right)$.

Corollary C.2. *Assume that the bound of the feature vectors* $D_{\boldsymbol{\varphi}}$ *is of order* $O(1)$*, that* $D_{\boldsymbol{\omega}} = D_{\boldsymbol{\psi}} = \sqrt{d}$ *which together imply* $D_{\boldsymbol{\theta}} \le \sqrt{d}+1+\sqrt{d}D_q = O(\sqrt{d}D_q)$ *and that* $D^2_{\boldsymbol{\beta}} = c \cdot C_{\varphi,c}(\pi^*;\pi_B)$ *for some positive universal constant* c*. Then, under the same assumptions of Theorem 3.2,* n_ε *is of order* $O\left(\varepsilon^{-4} D^4_q C_{\varphi,c}(\pi^*;\pi_B)^2 d^{4-2c} \log|\mathcal{A}|\right)$.

Recall that $C_{\varphi,1/2}$ is always smaller than $C_{\varphi,1}$, but using $c = 1/2$ in the algorithm requires knowledge of the covariance matrix $\boldsymbol{\Lambda}$, and results in a slightly worse dependence on the dimension.

The proof of Theorem C.1 mainly follows the same steps as in the discounted case, with some added difficulty that is inherent in the more challenging average-reward setup. Some key challenges include treating the additional optimization variable ρ and coping with the fact that the optimal parameters $\boldsymbol{\theta}^*$ and $\boldsymbol{\beta}^*$ are not necessarily unique any more.

C.2 Analysis

We now prove our main result regarding the AMDP setting in Theorem C.1. Following the derivations in the main text, we study the dynamic duality gap defined as

$$\mathcal{G}_T(\boldsymbol{\beta}^*, \pi^*; \rho^*_{1:T}, \boldsymbol{\theta}^*_{1:T}) = \frac{1}{T}\sum_{t=1}^{T}\big(f(\boldsymbol{\beta}^*, \pi^*; \rho_t, \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}_t, \pi_t; \rho^*_t, \boldsymbol{\theta}^*_t)\big). \tag{26}$$

First we show in Lemma C.3 below that, for appropriately chosen comparator points, the expected suboptimality of the policy returned by Algorithm 2 can be upper bounded in terms of the expected dynamic duality gap.

Lemma C.3. *Let* $\boldsymbol{\theta}^*_t$ *such that* $\langle \boldsymbol{\varphi}(x,a), \boldsymbol{\theta}^*_t\rangle = \langle \boldsymbol{\varphi}(x,a), \boldsymbol{\theta}^{\pi_t}\rangle - \inf_{(x,a)\in\mathcal{X}\times\mathcal{A}} \langle \boldsymbol{\varphi}(x,a), \boldsymbol{\theta}^{\pi_t}\rangle$ *holds for all* $(x,a) \in \mathcal{X}\times\mathcal{A}$*, and let* $\boldsymbol{v}^*_t$ *be defined as* $\boldsymbol{v}^*_t(x) = \sum_{a\in\mathcal{A}} \pi_t(a|x)\langle \boldsymbol{\varphi}(x,a), \boldsymbol{\theta}^*_t\rangle$ *for all* x*. Also, let* $\rho^*_t = \rho^{\pi_t}$*,* π^* *be an optimal policy, and* $\boldsymbol{\beta}^* = \boldsymbol{\Lambda}^{-c}\boldsymbol{\Phi}^\top\boldsymbol{\mu}^*$ *where* $\boldsymbol{\mu}^*$ *is the occupancy measure of* π^**. Then, the suboptimality gap of the policy output by Algorithm 2 satisfies*

$$\mathbb{E}_T\left[\langle \boldsymbol{\mu}^* - \boldsymbol{\mu}^{\boldsymbol{\pi}_{out}}, \boldsymbol{r}\rangle\right] = \mathcal{G}_T(\boldsymbol{\beta}^*, \pi^*; \rho^*_{1:T}, \boldsymbol{\theta}^*_{1:T}).$$

Proof. Substituting $(\boldsymbol{\beta}^*, \pi^*) = (\boldsymbol{\Lambda}^{-c}\boldsymbol{\Phi}^\intercal\boldsymbol{\mu}^*, \pi^*)$ in the first term of the dynamic duality gap we have

$$\begin{aligned} f(\boldsymbol{\beta}^*, \pi^*; \rho_t, \boldsymbol{\theta}_t) &= \rho_t + \langle \boldsymbol{\Lambda}^{-c}\boldsymbol{\Phi}^\intercal\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\theta}_t,\pi^*} - \boldsymbol{\theta}_t - \rho_t\boldsymbol{\varrho}]\rangle \\ &= \rho_t + \langle \boldsymbol{\mu}^*, r + \boldsymbol{P}\boldsymbol{v}_{\boldsymbol{\theta}_t,\pi^*} - \boldsymbol{\Phi}\boldsymbol{\theta}_t - \rho_t\mathbf{1}\rangle \\ &= \langle \boldsymbol{\mu}^*, r\rangle + \langle \boldsymbol{\mu}^*, \boldsymbol{E}\boldsymbol{v}_{\boldsymbol{\theta}_t,\pi^*} - \boldsymbol{\Phi}\boldsymbol{\theta}_t\rangle + \rho_t[1 - \langle \boldsymbol{\mu}^*, \mathbf{1}\rangle] \\ &= \langle \boldsymbol{\mu}^*, r\rangle. \end{aligned}$$

Here, we have used the fact that $\boldsymbol{\mu}^*$ is a valid occupancy measure, so it satisfies the flow constraint $\boldsymbol{E}^\intercal\boldsymbol{\mu}^* = \boldsymbol{P}^\intercal\boldsymbol{\mu}^*$ and the normalization constraint $\langle \boldsymbol{\mu}^*, \mathbf{1}\rangle = 1$. Also, in the last step we have used the definition of $\boldsymbol{v}_{\theta_t,\pi^*}$ that guarantees that the following equality holds:

$$\langle \boldsymbol{\mu}^*, \boldsymbol{\Phi}\boldsymbol{\theta}_t\rangle = \sum_{x\in\mathcal{X}} \nu^*(x) \sum_{a\in\mathcal{A}} \pi^*(a|x)\langle \boldsymbol{\theta}_t, \boldsymbol{\varphi}(x,a)\rangle = \sum_{x\in\mathcal{X}} \nu^*(x) v_{\boldsymbol{\theta}_t,\pi^*}(x) = \langle \boldsymbol{\mu}^*, \boldsymbol{E}\boldsymbol{v}_{\boldsymbol{\theta}_t,\pi^*}\rangle.$$

For the second term in the dynamic duality gap, using that π_t is $\mathcal{F}_{t-1}$-measurable we write

$$
\begin{aligned}
&f(\boldsymbol{\beta}_t, \pi_t; \rho_t^*, \boldsymbol{\theta}_t^*) \\
&\quad = \rho_t^* + \langle \boldsymbol{\beta}_t \,, \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\theta}_t^*, \pi_t} - \boldsymbol{\theta}_t^* - \rho_t^* \boldsymbol{\varrho}] \rangle \\
&\quad = \rho_t^* + \langle \boldsymbol{\beta}_t \,, \boldsymbol{\Lambda}^{c-1} \mathbb{E}_t \left[\boldsymbol{\varphi}_t \boldsymbol{\varphi}_t^\intercal [\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\theta}_t^*, \pi_t} - \boldsymbol{\theta}_t^* - \rho_t^* \boldsymbol{\varrho}] \right] \rangle \\
&\quad = \rho_t^* + \left\langle \boldsymbol{\beta}_t, \mathbb{E}_t \left[\boldsymbol{\Lambda}^{c-1} \boldsymbol{\varphi}_t \left[R_t + \sum_{x,a} p(x|X_t, A_t) \pi_t(a|x) \left\langle \boldsymbol{\varphi}(x,a), \boldsymbol{\theta}_t^* \right\rangle - \left\langle \boldsymbol{\varphi}(X_t, A_t), \boldsymbol{\theta}_t^* \right\rangle - \rho_t^* \right] \right] \right\rangle \\
&\quad = \rho^{\pi_t} + \left\langle \boldsymbol{\beta}_t, \mathbb{E}_t \left[\boldsymbol{\Lambda}^{c-1} \boldsymbol{\varphi}_t \left[R_t + \sum_{x,a} p(x|X_t, A_t) \pi_t(a|x) \left\langle \boldsymbol{\varphi}(x,a), \boldsymbol{\theta}^{\pi_t} \right\rangle - \left\langle \boldsymbol{\varphi}(X_t, A_t), \boldsymbol{\theta}^{\pi_t} \right\rangle - \rho^{\pi_t} \right] \right] \right\rangle \\
&\quad = \rho^{\pi_t} + \langle \boldsymbol{\beta}_t \,, \mathbb{E}_t \left[\boldsymbol{\Lambda}^{c-1} \boldsymbol{\varphi}_t [r(X_t, A_t) + \langle p(\cdot|X_t, A_t), v^{\pi_t} \rangle - q^{\pi_t}(X_t, A_t) - \rho^{\pi_t}] \right] \rangle \\
&\quad = \rho^{\pi_t} = \langle \boldsymbol{\mu}^{\pi_t}, r \rangle \,,
\end{aligned}
$$

where in the fourth equality we used that $\langle \boldsymbol{\varphi}(x,a) - \boldsymbol{\varphi}(x',a'), \theta_t^* \rangle = \langle \boldsymbol{\varphi}(x,a) - \boldsymbol{\varphi}(x',a'), \theta^{\pi_t} \rangle$ holds for all x, a, x', a' by definition of θ_t^*. Then, the last equality follows from the fact that the Bellman equations for π_t imply $q^{\pi_t}(x,a) + \rho^{\pi_t} = r(x,a) + \langle p(\cdot|x,a), \boldsymbol{v}^{\pi_t} \rangle$.

Combining both expressions for $f(\boldsymbol{\beta}^*, \pi^*; \rho_t, \boldsymbol{\theta}_t)$ and $f(\boldsymbol{\beta}_t, \pi_t; \rho_t^*, \boldsymbol{\theta}_t^*)$ in the dynamic duality gap we have:

$$
\mathcal{G}_T(\boldsymbol{\beta}^*, \pi^*; \rho_{1:T}^*, \boldsymbol{\theta}_{1:T}^*) = \frac{1}{T} \sum_{t=1}^{T} \left(\langle \boldsymbol{\mu}^* - \boldsymbol{\mu}^{\pi_t}, r \rangle \right) = \mathbb{E}_T \left[\langle \boldsymbol{\mu}^* - \boldsymbol{\mu}^{\pi_{\text{out}}}, r \rangle \right].
$$

The second equality follows from noticing that, since $\boldsymbol{\pi}_{\text{out}}$ is sampled uniformly from $\{\pi_t\}_{t=1}^T$, $\mathbb{E}\left[\langle \boldsymbol{\mu}^{\boldsymbol{\pi}_{\text{out}}}, \boldsymbol{r} \rangle\right] = \frac{1}{T} \sum_{t=1}^T \mathbb{E}\left[\langle \boldsymbol{\mu}^{\pi_t}, \boldsymbol{r} \rangle\right]$. This completes the proof. □

Having shown that for well-chosen comparator points the dynamic duality gap equals the expected suboptimality of the output policy of Algorithm 2, it remains to relate the gap to the optimization error of the primal-dual procedure. This is achieved in the following lemma.

Lemma C.4. *For the same choice of comparators $(\boldsymbol{\beta}^*, \pi^*; \rho_{1:T}^*, \boldsymbol{\theta}_{1:T}^*)$ as in Lemma C.3 the dynamic duality gap associated with the iterates produced by Algorithm 2 satisfies*

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{G}_T(\boldsymbol{\beta}^*, \pi^*; \rho_{1:T}^*, \boldsymbol{\theta}_{1:T}^*)\right] \\
&\qquad \leq \frac{2D_{\boldsymbol{\beta}}^2}{\zeta T} + \frac{\mathcal{H}\left(\pi^* \| \pi_1\right)}{\alpha T} + \frac{1}{2\xi K} + \frac{2D_{\boldsymbol{\theta}}^2}{\eta K} \\
&\qquad\quad + \frac{\zeta \operatorname{Tr}(\boldsymbol{\Lambda}^{2c-1})(1 + 2D_{\boldsymbol{\varphi}} D_{\boldsymbol{\theta}})^2}{2} + \frac{\alpha D_{\boldsymbol{\varphi}}^2 D_{\boldsymbol{\theta}}^2}{2} + \xi \left(1 + D_{\boldsymbol{\beta}}^2 \left\|\boldsymbol{\Lambda}\right\|_2^{2c-1}\right) + 2\eta D_{\boldsymbol{\varphi}}^2 D_{\boldsymbol{\beta}}^2 \left\|\boldsymbol{\Lambda}\right\|_2^{2c-1}.
\end{aligned}
$$

Proof. The first part of the proof follows from recognising that the dynamic duality gap can be rewritten in terms of the total regret of the primal and dual players in the algorithm. Formally, we write

$$
\begin{aligned}
&\mathcal{G}_T(\boldsymbol{\beta}^*, \pi^*; \rho_{1:T}^*, \boldsymbol{\theta}_{1:T}^*) \\
&\qquad = \frac{1}{T} \sum_{t=1}^{T} \left(f(\boldsymbol{\beta}^*, \pi^*; \rho_t, \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}_t, \pi_t; \rho_t, \boldsymbol{\theta}_t) \right) + \frac{1}{T} \sum_{t=1}^{T} \left(f(\boldsymbol{\beta}_t, \pi_t; \rho_t, \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}_t, \pi_t; \rho_t^*, \boldsymbol{\theta}_t^*) \right).
\end{aligned}
$$

Using that $\boldsymbol{\beta}^* = \boldsymbol{\Lambda}^{-c} \boldsymbol{\Phi}^\top \boldsymbol{\mu}^*$, $\boldsymbol{q}_t = \langle \boldsymbol{\varphi}(x,a), \boldsymbol{\theta}_t \rangle$, $\boldsymbol{v}_t = \boldsymbol{v}_{\boldsymbol{\theta}_t, \pi_t}$ and that $\boldsymbol{g}_{\boldsymbol{\beta},t} = \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_t - \boldsymbol{\theta}_t - \rho_t \boldsymbol{\varrho}]$, we see that term in the first sum can be simply rewritten as

$$
\begin{aligned}
&f(\boldsymbol{\beta}^*, \pi^*; \rho_t, \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}_t, \pi_t; \rho_t, \boldsymbol{\theta}_t) \\
&\qquad = \langle \boldsymbol{\beta}^* \,, \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\theta}_t, \pi^*} - \boldsymbol{\theta}_t - \rho_t \boldsymbol{\varrho}] \rangle - \langle \boldsymbol{\beta}_t \,, \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\theta}_t, \pi_t} - \boldsymbol{\theta}_t - \rho_t \boldsymbol{\varrho}] \rangle \\
&\qquad = \langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_t \,, \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_t - \boldsymbol{\theta}_t - \rho_t \boldsymbol{\varrho}] \rangle + \langle \boldsymbol{\Psi}^\intercal \boldsymbol{\Lambda}^c \boldsymbol{\beta}^* \,, \boldsymbol{v}_{\boldsymbol{\theta}_t, \pi^*} - \boldsymbol{v}_{\boldsymbol{\theta}_t, \pi_t} \rangle \\
&\qquad = \langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_t \,, \boldsymbol{g}_{\boldsymbol{\beta},t} \rangle + \sum_{x \in \mathcal{X}} \nu^*(x) \left\langle \pi^*(\cdot|x) - \pi_t(\cdot|x), \boldsymbol{q}_t(x, \cdot) \right\rangle.
\end{aligned}
$$

In a similar way, using that $\boldsymbol{E}^\intercal \boldsymbol{\mu}_t = \boldsymbol{\Psi}^\intercal \boldsymbol{\Lambda}^c \boldsymbol{\beta}_t$ and the definitions of the gradients $g_{\rho,t}$ and $\boldsymbol{g}_{\boldsymbol{\theta},t}$, the term in the second sum can be rewritten as

$$
\begin{aligned}
&f(\boldsymbol{\beta}_t, \pi_t; \rho_t, \boldsymbol{\theta}_t) - f(\boldsymbol{\beta}_t, \pi_t; \rho_t^*, \boldsymbol{\theta}_t^*) \\
&\quad = \rho_t + \langle \boldsymbol{\beta}_t \,, \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\theta}_t, \pi_t} - \boldsymbol{\theta}_t - \rho_t \boldsymbol{\varrho}] \rangle - \rho_t^* - \langle \boldsymbol{\beta}_t \,, \boldsymbol{\Lambda}^c[\boldsymbol{\omega} + \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\theta}_t^*, \pi_t} - \boldsymbol{\theta}_t^* - \rho_t^* \boldsymbol{\varrho}] \rangle \\
&\quad = (\rho_t - \rho_t^*)[1 - \langle \boldsymbol{\beta}_t, \boldsymbol{\Lambda}^c \boldsymbol{\varrho} \rangle] - \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*, \boldsymbol{\Lambda}^c \boldsymbol{\beta}_t \rangle + \left\langle \boldsymbol{E}^\intercal \boldsymbol{\mu}_t, \boldsymbol{v}_{\boldsymbol{\theta}_t, \pi_t} - \boldsymbol{v}_{\boldsymbol{\theta}_t^*, \pi_t} \right\rangle \\
&\quad = (\rho_t - \rho_t^*)[1 - \langle \boldsymbol{\beta}_t, \boldsymbol{\Lambda}^c \boldsymbol{\varrho} \rangle] - \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*, \boldsymbol{\Lambda}^c \boldsymbol{\beta}_t \rangle + \langle \boldsymbol{\Phi}^\intercal \boldsymbol{\mu}_t, \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^* \rangle \\
&\quad = (\rho_t - \rho_t^*)[1 - \langle \boldsymbol{\beta}_t, \boldsymbol{\Lambda}^c \boldsymbol{\varrho} \rangle] + \langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*, \boldsymbol{\Phi}^\intercal \boldsymbol{\mu}_t - \boldsymbol{\Lambda}^c \boldsymbol{\beta}_t \rangle \\
&\quad = (\rho_t - \rho_t^*) g_{\rho,t} + \left\langle \boldsymbol{\theta}_t - \boldsymbol{\theta}_t^*, \boldsymbol{g}_{\boldsymbol{\theta},t} \right\rangle = \frac{1}{K} \sum_{i=1}^{K} \left((\rho_t^{(i)} - \rho_t^*) g_{\rho,t} + \left\langle \boldsymbol{\theta}_t^{(i)} - \boldsymbol{\theta}_t^*, \boldsymbol{g}_{\boldsymbol{\theta},t} \right\rangle \right).
\end{aligned}
$$

Combining both terms in the duality gap concludes the first part of the proof. As shown below the dynamic duality gap is written as the error between iterates of the algorithm from respective comparator points in the direction of the exact gradients. Formally, we have

$$
\begin{aligned}
\mathcal{G}_T(\boldsymbol{\beta}^*, \pi^*; \rho_{1:T}^*, \boldsymbol{\theta}_{1:T}^*) = \frac{1}{T} \sum_{t=1}^{T} \Bigg(&\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_t \,, \boldsymbol{g}_{\boldsymbol{\beta},t} \rangle + \sum_{x \in \mathcal{X}} \nu^*(x) \, \langle \pi^*(\cdot|x) - \pi_t(\cdot|x), \boldsymbol{q}_t(x, \cdot) \rangle \Bigg) \\
&+ \frac{1}{TK} \sum_{t=1}^{T} \sum_{i=1}^{K} \left((\rho_t^{(i)} - \rho_t^*) g_{\rho,t} + \left\langle \boldsymbol{\theta}_t^{(i)} - \boldsymbol{\theta}_t^*, \boldsymbol{g}_{\boldsymbol{\theta},t} \right\rangle \right).
\end{aligned}
$$

Then, implementing techniques from stochastic gradient descent analysis in the proof of Lemmas C.5 to C.7 and mirror descent analysis in Lemma B.3, the expected dynamic duality gap can be upper bounded as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{G}_T(\boldsymbol{\beta}^*, \pi^*; \rho_{1:T}^*, \boldsymbol{\theta}_{1:T}^*)\right] \\
&\quad \leq \frac{2D_{\boldsymbol{\beta}}^2}{\zeta T} + \frac{\mathcal{H}\left(\pi^* \| \pi_1\right)}{\alpha T} + \frac{1}{2\xi K} + \frac{2D_{\boldsymbol{\theta}}^2}{\eta K} \\
&\qquad + \frac{\zeta \operatorname{Tr}(\boldsymbol{\Lambda}^{2c-1})(1 + 2D_{\boldsymbol{\varphi}} D_{\boldsymbol{\theta}})^2}{2} + \frac{\alpha D_{\boldsymbol{\varphi}}^2 D_{\boldsymbol{\theta}}^2}{2} + \xi \left(1 + D_{\boldsymbol{\beta}}^2 \left\|\boldsymbol{\Lambda}\right\|_2^{2c-1}\right) + 2\eta D_{\boldsymbol{\varphi}}^2 D_{\boldsymbol{\beta}}^2 \left\|\boldsymbol{\Lambda}\right\|_2^{2c-1}.
\end{aligned}
$$

This completes the proof □

Proof of Theorem C.1 First, we bound the expected suboptimality gap by combining Lemma C.3 and C.4. Next, bearing in mind that the algorithm only needs $T(K+1)$ total samples from the behavior policy we optimize the learning rates to obtain a bound on the sample complexity, thus completing the proof. □

C.3 Missing proofs for Lemma C.4

In this section we prove Lemmas C.5 to C.7 used in the proof of Lemma C.4. It is important to recall that sample transitions (X_k, A_k, R_t, X_k') in any iteration k are generated in the following way: we draw i.i.d state-action pairs (X_k, A_k) from $\boldsymbol{\mu}_B$, and for each state-action pair, the next X_k' is sampled from $p(\cdot|X_k, A_k)$ and immediate reward computed as $R_t = r(X_k, A_k)$. Precisely in iteration i of round t where $k = (t, i)$, since $(X_{t,i}, A_{t,i})$ are sampled i.i.d from $\boldsymbol{\mu}_B$ at this time step, $\mathbb{E}_{t,i}\left[\boldsymbol{\varphi}_{t,i} \boldsymbol{\varphi}_{t,i}^\intercal\right] = \mathbb{E}_{(x,a)\sim\boldsymbol{\mu}_B}\left[\boldsymbol{\varphi}(x,a)\boldsymbol{\varphi}(x,a)^\intercal\right] = \boldsymbol{\Lambda}$.

Lemma C.5. *The gradient estimator $\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t}$ satisfies $\mathbb{E}\left[\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t} \,|\mathcal{F}_{t-1}, \boldsymbol{\theta}_t\right] = \boldsymbol{g}_{\boldsymbol{\beta},t}$ and*

$$
\mathbb{E}\left[\|\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t}\|_2^2\right] \leq \operatorname{Tr}(\boldsymbol{\Lambda}^{2c-1})(1 + 2D_{\boldsymbol{\varphi}} D_{\boldsymbol{\theta}})^2.
$$

Furthermore, for any $\boldsymbol{\beta}^*$ with $\boldsymbol{\beta}^* \in \mathbb{B}(D_{\boldsymbol{\beta}})$, the iterates $\boldsymbol{\beta}_t$ satisfy

$$
\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{\beta}^* - \boldsymbol{\beta}_t \,, \boldsymbol{g}_{\boldsymbol{\beta},t} \rangle\right] \leq \frac{2D_{\boldsymbol{\beta}}^2}{\zeta} + \frac{\zeta T \operatorname{Tr}(\boldsymbol{\Lambda}^{2c-1})(1 + 2D_{\boldsymbol{\varphi}} D_{\boldsymbol{\theta}})^2}{2}. \tag{27}
$$

Proof. For the first part, we remind that π_t is $\mathcal{F}_{t-1}$-measurable and $\boldsymbol{v}_t$ is determined given π_t and $\boldsymbol{\theta}_t$. Then, we write

$$\begin{aligned}
\mathbb{E}\left[\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t}\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right] &= \mathbb{E}\left[\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_t[R_t + v_t(X_t') - \langle\boldsymbol{\theta}_t,\boldsymbol{\varphi}_t\rangle - \rho_t]\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right]\\
&= \mathbb{E}\left[\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_t[R_t + \mathbb{E}_{x'\sim p(\cdot|X_t,A_t)}\left[v_t(x')\right] - \langle\boldsymbol{\theta}_t,\boldsymbol{\varphi}_t\rangle - \rho_t]\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right]\\
&= \mathbb{E}\left[\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_t[R_t + \langle p(\cdot|X_t,A_t),\boldsymbol{v}_t\rangle - \langle\boldsymbol{\theta}_t,\boldsymbol{\varphi}_t\rangle - \rho_t]\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right]\\
&= \mathbb{E}\left[\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_t\boldsymbol{\varphi}_t^\intercal[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_t - \boldsymbol{\theta}_t - \rho_t\boldsymbol{\varrho}]\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right]\\
&= \boldsymbol{\Lambda}^{c-1}\mathbb{E}\left[\boldsymbol{\varphi}_t\boldsymbol{\varphi}_t^\intercal\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right][\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_t - \boldsymbol{\theta}_t - \rho_t\boldsymbol{\varrho}]\\
&= \boldsymbol{\Lambda}^{c}[\boldsymbol{\omega} + \boldsymbol{\Psi}\boldsymbol{v}_t - \boldsymbol{\theta}_t - \rho_t\boldsymbol{\varrho}] = \boldsymbol{g}_{\boldsymbol{\beta},t}.
\end{aligned}$$

Next, we use the facts that $r\in[0,1]$ and $\|\boldsymbol{v}_t\|_\infty \le \|\boldsymbol{\Phi}\boldsymbol{\theta}_t\|_\infty \le D_{\boldsymbol{\varphi}}D_{\boldsymbol{\theta}}$ to show the following bound:

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\boldsymbol{g}}_{\boldsymbol{\beta},t}\|_2^2\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right] &= \mathbb{E}\left[\left\|\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_t[R_t + v_t(X_t') - \langle\boldsymbol{\theta}_t,\boldsymbol{\varphi}_t\rangle]\right\|_2^2\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right]\\
&= \mathbb{E}\left[|R_t + v_t(X_t') - \langle\boldsymbol{\theta}_t,\boldsymbol{\varphi}_t\rangle|\left\|\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_t\right\|_2^2\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right]\\
&\le \mathbb{E}\left[(1+2D_{\boldsymbol{\varphi}}D_{\boldsymbol{\theta}})^2\left\|\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_t\right\|_2^2\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right]\\
&= (1+2D_{\boldsymbol{\varphi}}D_{\boldsymbol{\theta}})^2\mathbb{E}\left[\boldsymbol{\varphi}_t^\intercal\boldsymbol{\Lambda}^{2(c-1)}\boldsymbol{\varphi}_t\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right]\\
&= (1+2D_{\boldsymbol{\varphi}}D_{\boldsymbol{\theta}})^2\mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\Lambda}^{2(c-1)}\boldsymbol{\varphi}_t\boldsymbol{\varphi}_t^\intercal)\,|\mathcal{F}_{t-1},\boldsymbol{\theta}_t\right]\\
&\le \mathrm{Tr}(\boldsymbol{\Lambda}^{2c-1})(1+2D_{\boldsymbol{\varphi}}D_{\boldsymbol{\theta}})^2.
\end{aligned}$$

The last step follows from the fact that $\boldsymbol{\Lambda}$, hence also $\boldsymbol{\Lambda}^{2c-1}$, is positive semi-definite, so $\mathrm{Tr}(\boldsymbol{\Lambda}^{2c-1})\ge 0$. Having shown these properties, we appeal to the standard analysis of online gradient descent stated as Lemma D.1 to obtain the following bound

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle\boldsymbol{\beta}^* - \boldsymbol{\beta}_t\,,\boldsymbol{g}_{\boldsymbol{\beta},t}\rangle\right] \le \frac{\|\boldsymbol{\beta}_1-\boldsymbol{\beta}^*\|_2^2}{2\zeta} + \frac{\zeta T\,\mathrm{Tr}(\boldsymbol{\Lambda}^{2c-1})(1+2D_{\boldsymbol{\varphi}}D_{\boldsymbol{\theta}})^2}{2}.$$

Using that $\|\boldsymbol{\beta}^*\|_2 \le D_{\boldsymbol{\beta}}$ concludes the proof. □

Lemma C.6. *The gradient estimator $\tilde{g}_{\rho,t,i}$ satisfies $\mathbb{E}_{t,i}\left[\tilde{g}_{\rho,t,i}\right] = g_{\rho,t}$ and $\mathbb{E}_{t,i}\left[\tilde{g}_{\rho,t,i}^2\right] \le 2 + 2D_{\boldsymbol{\beta}}^2\|\boldsymbol{\Lambda}\|_2^{2c-1}$. Furthermore, for any $\rho_t^*\in[0,1]$, the iterates $\rho_t^{(i)}$ satisfy*

$$\mathbb{E}\left[\sum_{i=1}^{K}(\rho_t^{(i)} - \rho_t^*)g_{\rho,t}\right] \le \frac{1}{2\xi} + \xi K\left(1 + \|\boldsymbol{\beta}_t\|_{\boldsymbol{\Lambda}^{2c-1}}^2\right).$$

Proof. For the first part of the proof, we use that $\boldsymbol{\beta}_t$ is $\mathcal{F}_{t,i-1}$-measurable, to obtain

$$\begin{aligned}
\mathbb{E}_{t,i}\left[\tilde{g}_{\rho,t,i}\right] &= \mathbb{E}_{t,i}\left[1 - \left\langle\boldsymbol{\varphi}_{t,i},\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right\rangle\right]\\
&= \mathbb{E}_{t,i}\left[1 - \left\langle\boldsymbol{\varphi}_{t,i}\boldsymbol{\varphi}_{t,i}^\intercal\boldsymbol{\varrho},\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right\rangle\right]\\
&= 1 - \langle\boldsymbol{\Lambda}^c\boldsymbol{\varrho},\boldsymbol{\beta}_t\rangle = g_{\rho,t}.
\end{aligned}$$

In addition, using Young's inequality and $\|\boldsymbol{\beta}_t\|_{\boldsymbol{\Lambda}^{2c-1}}^2 \le D_{\boldsymbol{\beta}}^2\|\boldsymbol{\Lambda}\|_2^{2c-1}$ we show that

$$\begin{aligned}
\mathbb{E}_{t,i}\left[\tilde{g}_{\rho,t,i}^2\right] &= \mathbb{E}_{t,i}\left[\left(1 - \left\langle\boldsymbol{\varphi}_{t,i},\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right\rangle\right)^2\right]\\
&\le 2 + 2\mathbb{E}_{t,i}\left[\boldsymbol{\beta}_t^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_{t,i}\boldsymbol{\varphi}_{t,i}^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right]\\
&= 2 + 2\|\boldsymbol{\beta}_t\|_{\boldsymbol{\Lambda}^{2c-1}}^2 \le 2 + 2D_{\boldsymbol{\beta}}^2\|\boldsymbol{\Lambda}\|_2^{2c-1}.
\end{aligned}$$

For the second part, we appeal to the standard online gradient descent analysis of Lemma D.1 to bound on the total error of the iterates:

$$\mathbb{E}\left[\sum_{i=1}^{K}(\rho_t^{(i)} - \rho_t^*)g_{\rho,t}\right] \le \frac{\left(\rho_t^{(1)} - \rho_t^*\right)^2}{2\xi} + \xi K\left(1 + D_{\boldsymbol{\beta}}^2\|\boldsymbol{\Lambda}\|_2^{2c-1}\right).$$

Using that $\left(\rho_t^{(1)} - \rho_t^*\right)^2 \leq 1$ concludes the proof. □

Lemma C.7. *The gradient estimator $\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,i}$ satisfies $\mathbb{E}_{t,i}\left[\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,i}\right] = \boldsymbol{g}_{\boldsymbol{\theta},t,i}$ and $\mathbb{E}_{t,i}\left[\|\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,i}\|_2^2\right] \leq 4D_{\boldsymbol{\varphi}}^2 D_{\boldsymbol{\beta}}^2 \|\boldsymbol{\Lambda}\|_2^{2c-1}$. Furthermore, for any $\boldsymbol{\theta}_t^*$ with $\|\boldsymbol{\theta}_t^*\|_2 \leq D_{\boldsymbol{\theta}}$, the iterates $\boldsymbol{\theta}_t^{(i)}$ satisfy*

$$\mathbb{E}\left[\sum_{i=1}^{K}\left\langle \boldsymbol{\theta}_t^{(i)} - \boldsymbol{\theta}_t^*, \boldsymbol{g}_{\boldsymbol{\theta},t,i}\right\rangle\right] \leq \frac{2D_{\boldsymbol{\theta}}^2}{\eta} + 2\eta K D_{\boldsymbol{\varphi}}^2 D_{\boldsymbol{\beta}}^2 \|\boldsymbol{\Lambda}\|_2^{2c-1}. \tag{28}$$

Proof. Since $\boldsymbol{\beta}_t, \pi_t, \rho_t^i$ and $\boldsymbol{\theta}_t^i$ are $\mathcal{F}_{t,i-1}$-measurable, we obtain

$$\begin{aligned}
\mathbb{E}_{t,i}\left[\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,i}\right] &= \mathbb{E}_{t,i}\left[\boldsymbol{\varphi}'_{t,i}\left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right\rangle - \boldsymbol{\varphi}_{t,i}\left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right\rangle\right] \\
&= \boldsymbol{\Phi}^\intercal \mathbb{E}_{t,i}\left[\boldsymbol{e}_{X'_{t,i},A'_{t,i}}\left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right\rangle\right] - \mathbb{E}_{t,i}\left[\boldsymbol{\varphi}_{t,i}\boldsymbol{\varphi}_{t,i}^\intercal\right]\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t \\
&= \boldsymbol{\Phi}^\intercal \mathbb{E}_{t,i}\left[[\pi_t \circ p(\cdot|X_t, A_t)]\left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right\rangle\right] - \boldsymbol{\Lambda}^c\boldsymbol{\beta}_t \\
&= \boldsymbol{\Phi}[\pi_t \circ \boldsymbol{\Psi}^\intercal \mathbb{E}_{t,i}\left[\boldsymbol{\varphi}_{t,i}\boldsymbol{\varphi}_{t,i}^\intercal\right]\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t] - \boldsymbol{\Lambda}^c\boldsymbol{\beta}_t \\
&= \boldsymbol{\Phi}[\pi_t \circ \boldsymbol{\Psi}^\intercal\boldsymbol{\Lambda}^c\boldsymbol{\beta}_t] - \boldsymbol{\Lambda}^c\boldsymbol{\beta}_t \\
&= \boldsymbol{\Phi}^\intercal\boldsymbol{\mu}_t - \boldsymbol{\Lambda}^c\boldsymbol{\beta}_t = \boldsymbol{g}_{\boldsymbol{\theta},t}.
\end{aligned}$$

Next, we consider the squared gradient norm and bound it via elementary manipulations as follows:

$$\begin{aligned}
\mathbb{E}_{t,i}\left[\left\|\tilde{\boldsymbol{g}}_{\boldsymbol{\theta},t,i}\right\|_2^2\right] &= \mathbb{E}_{t,i}\left[\left\|\boldsymbol{\varphi}'_{t,i}\left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right\rangle - \boldsymbol{\varphi}_{t,i}\left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right\rangle\right\|_2^2\right] \\
&\leq 2\mathbb{E}_{t,i}\left[\left\|\boldsymbol{\varphi}'_{t,i}\left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right\rangle\right\|_2^2\right] + 2\mathbb{E}_{t,i}\left[\left\|\boldsymbol{\varphi}_{t,i}\left\langle \boldsymbol{\varphi}_{t,i}, \boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right\rangle\right\|_2^2\right] \\
&= 2\mathbb{E}_{t,i}\left[\boldsymbol{\beta}_t^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_{t,i}\left\|\boldsymbol{\varphi}'_{t,i}\right\|_2^2\boldsymbol{\varphi}_{t,i}^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right] + 2\mathbb{E}_{t,i}\left[\boldsymbol{\beta}_t^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_{t,i}\|\boldsymbol{\varphi}_{t,i}\|_2^2\boldsymbol{\varphi}_{t,i}^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right] \\
&\leq 2D_{\boldsymbol{\varphi}}^2\mathbb{E}_{t,i}\left[\boldsymbol{\beta}_t^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_{t,i}\boldsymbol{\varphi}_{t,i}^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right] + 2D_{\boldsymbol{\varphi}}^2\mathbb{E}_{t,i}\left[\boldsymbol{\beta}_t^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\varphi}_{t,i}\boldsymbol{\varphi}_{t,i}^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right] \\
&= 2D_{\boldsymbol{\varphi}}^2\mathbb{E}_{t,i}\left[\boldsymbol{\beta}_t^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right] + 2D_{\boldsymbol{\varphi}}^2\mathbb{E}_{t,i}\left[\boldsymbol{\beta}_t^\intercal\boldsymbol{\Lambda}^{c-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{c-1}\boldsymbol{\beta}_t\right] \\
&\leq 4D_{\boldsymbol{\varphi}}^2\|\boldsymbol{\beta}_t\|_{\boldsymbol{\Lambda}^{2c-1}}^2 \leq 4D_{\boldsymbol{\varphi}}^2 D_{\boldsymbol{\beta}}^2\|\boldsymbol{\Lambda}\|_2^{2c-1}.
\end{aligned}$$

Having verified these conditions, we appeal to the online gradient descent analysis of Lemma D.1 to show the bound

$$\mathbb{E}\left[\sum_{i=1}^{K}\left\langle \boldsymbol{\theta}_t^{(i)} - \boldsymbol{\theta}_t^*, \boldsymbol{g}_{\boldsymbol{\theta},t}\right\rangle\right] \leq \frac{\left\|\boldsymbol{\theta}_t^{(1)} - \boldsymbol{\theta}_t^*\right\|_2^2}{2\eta} + 2\eta K D_{\boldsymbol{\varphi}}^2 D_{\boldsymbol{\beta}}^2\|\boldsymbol{\Lambda}\|_2^{2c-1}.$$

We then use that $\left\|\boldsymbol{\theta}_t^* - \boldsymbol{\theta}_t^{(1)}\right\|_2 \leq 2D_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_t^*, \boldsymbol{\theta}_t^{(1)} \in \mathbb{B}(D_{\boldsymbol{\theta}})$, thus concluding the proof. □

D AUXILIARY LEMMAS

The following is a standard result in convex optimization proved here for the sake of completeness—we refer to Nemirovski and Yudin (1983); Zinkevich (2003); Orabona (2019) for more details and comments on the history of this result.

Lemma D.1 (Online Stochastic Gradient Descent)**.** *Given $y_1 \in \mathbb{B}(D_y)$ and $\eta > 0$, define the sequences $y_2, \cdots, y_{n+1}$ and $h_1, \cdots, h_n$ such that for $k = 1, \cdots, n$,*

$$y_{k+1} = \Pi_{\mathbb{B}(D_y)} \left(y_k + \eta \widehat{h}_k \right),$$

and $\widehat{h}_k$ satisfies $\mathbb{E}\left[\widehat{h}_k \,|\mathcal{F}_{k-1}\right] = h_k$ and $\mathbb{E}\left[\left\|\widehat{h}_k\right\|_2^2 |\mathcal{F}_{k-1}\right] \leq G^2$. Then, for $y^* \in \mathbb{B}(D_y)$:

$$\mathbb{E}\left[\sum_{k=1}^{n} \langle y^* - y_k, h_k \rangle\right] \leq \frac{\|y_1 - y^*\|_2^2}{2\eta} + \frac{\eta n G^2}{2}.$$

Proof. We start by studying the following term:

$$\begin{aligned}
\|y_{k+1} - y^*\|_2^2 &= \left\|\Pi_{\mathbb{B}(D_y)}(y_k + \eta \widehat{h}_k) - y^*\right\|_2^2 \\
&\leq \left\|y_k + \eta \widehat{h}_k - y^*\right\|_2^2 \\
&= \|y_k - y^*\|_2^2 - 2\eta \left\langle y^* - y_k, \widehat{h}_k \right\rangle + \eta^2 \left\|\widehat{h}_k\right\|_2^2.
\end{aligned}$$

The inequality is due to the fact that the projection operator is a non-expansion with respect to the Euclidean norm. Since $\mathbb{E}\left[\widehat{h}_k \,|\mathcal{F}_{k-1}\right] = h_k$, we can rearrange the above equation and take a conditional expectation to obtain

$$\begin{aligned}
\langle y^* - y_k, h_k \rangle &\leq \frac{\|y_k - y^*\|_2^2 - \mathbb{E}\left[\|y_{k+1} - y^*\|_2^2 \,|\mathcal{F}_{k-1}\right]}{2\eta} + \frac{\eta}{2}\mathbb{E}\left[\left\|\widehat{h}_k\right\|_2^2 |\mathcal{F}_{k-1}\right] \\
&\leq \frac{\|y_k - y^*\|_2^2 - \mathbb{E}\left[\|y_{k+1} - y^*\|_2^2 \,|\mathcal{F}_{k-1}\right]}{2\eta} + \frac{\eta G^2}{2},
\end{aligned}$$

where the last inequality is from $\mathbb{E}\left[\left\|\widehat{h}_k\right\|_2^2 |\mathcal{F}_{k-1}\right] \leq G^2$. Finally, taking a sum over $k = 1, \cdots, n$, taking a marginal expectation, evaluating the resulting telescoping sum and upper-bounding negative terms by zero we obtain the desired result as

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{n} \left\langle y^* - y_k, \hat{h}_k \right\rangle\right] &\leq \frac{\|y_1 - y^*\|_2^2 - \mathbb{E}\left[\|y_{n+1} - y^*\|_2^2\right]}{2\eta} + \frac{\eta}{2}\sum_{k=1}^{n} G^2 \\
&\leq \frac{\|y_1 - y^*\|_2^2}{2\eta} + \frac{\eta n G^2}{2}.
\end{aligned}$$

□

The next result is a similar regret analysis for mirror descent with the relative entropy as its distance generating function. Once again, this result is standard, and we refer the interested reader to Nemirovski and Yudin (1983); Cesa-Bianchi and Lugosi (2006); Orabona (2019) for more details. For the analysis, we recall that $\mathcal{D}$ denotes the relative entropy (or Kullback–Leibler divergence), defined for any $p, q \in \Delta_{\mathcal{A}}$ as $\mathcal{D}(p\|q) = \sum_a p(a) \log \frac{p(a)}{q(a)}$, and that, for any two policies π, π', we define the conditional entropy[4] $\mathcal{H}(\pi\|\pi') \doteq \sum_{x \in \mathcal{X}} \nu^{\pi}(x) \mathcal{D}(\pi(\cdot|x)\|\pi'(\cdot|x))$.

[4]Technically speaking, this quantity is the conditional entropy between the occupancy measures μ^{π} and $\mu^{\pi'}$. We will continue to use this relatively imprecise terminology to keep our notation light, and we refer to Neu et al. (2017) and Bas-Serrano et al. (2021) for more details.

Lemma D.2 (Mirror Descent)**.** *Let $q_t, \dots, q_T$ be a sequence of functions from $\mathcal{X} \times \mathcal{A}$ to $\mathbb{R}$ so that $\|q_t\|_\infty \leq D_q$ for $t = 1, \dots, T$. Given an initial policy π_1 and a learning rate $\alpha > 0$, define the sequence of policies $\pi_2, \dots, \pi_{T+1}$ such that, for $t = 1, \dots, T$:*

$$\pi_{t+1}(a|x) \propto \pi_t e^{\alpha q_t(x,a)}.$$

Then, for any comparator policy π^*:

$$\sum_{t=1}^{T} \sum_{x \in \mathcal{X}} \nu^{\pi^*}(x) \left\langle \pi^*(\cdot|x) - \pi_t(\cdot|x), q_t(x, \cdot) \right\rangle \leq \frac{\mathcal{H}\left(\pi^* \| \pi_1\right)}{\alpha} + \frac{\alpha T D_q^2}{2}.$$

Proof. We begin by studying the relative entropy between $\pi^*(\cdot|x)$ and iterates $\pi_t(\cdot|x), \pi_{t+1}(\cdot|x)$ for any $x \in \mathcal{X}$:

$$\begin{aligned}
\mathcal{D}\left(\pi^*(\cdot|x) \| \pi_{t+1}(\cdot|x)\right) &= \mathcal{D}\left(\pi^*(\cdot|x) \| \pi_t(\cdot|x)\right) - \sum_{a \in \mathcal{A}} \pi^*(a|x) \log \frac{\pi_{t+1}(a|x)}{\pi_t(a|x)} \\
&= \mathcal{D}\left(\pi^*(\cdot|x) \| \pi_t(\cdot|x)\right) - \sum_{a \in \mathcal{A}} \pi^*(a|x) \log \frac{e^{\alpha q_t(x,a)}}{\sum_{a' \in \mathcal{A}} \pi_t(a'|x) e^{\alpha q_t(x,a')}} \\
&= \mathcal{D}\left(\pi^*(\cdot|x) \| \pi_t(\cdot|x)\right) - \alpha \left\langle \pi^*(\cdot|x), q_t(x, \cdot) \right\rangle + \log \sum_{a \in \mathcal{A}} \pi_t(a|x) e^{\alpha q_t(x,a)} \\
&= \mathcal{D}\left(\pi^*(\cdot|x) \| \pi_t(\cdot|x)\right) - \alpha \left\langle \pi^*(\cdot|x) - \pi_t(\cdot|x), q_t(x, \cdot) \right\rangle \\
&\quad + \log \sum_{a \in \mathcal{A}} \pi_t(a|x) e^{\alpha q_t(x,a)} - \alpha \sum_{a \in \mathcal{A}} \pi_t(a|x) q_t(x, a) \\
&\leq \mathcal{D}\left(\pi^*(\cdot|x) \| \pi_t(\cdot|x)\right) - \alpha \left\langle \pi^*(\cdot|x) - \pi_t(\cdot|x), q_t(x, \cdot) \right\rangle + \frac{\alpha^2 \left\| q_t(x, \cdot) \right\|_\infty^2}{2}
\end{aligned}$$

where the last inequality follows from Hoeffding's lemma (cf. Lemma A.1 in Cesa-Bianchi and Lugosi, 2006). Next, we rearrange the above equation, sum over $t = 1, \cdots, T$, evaluate the resulting telescoping sum and upper-bound negative terms by zero to obtain

$$\sum_{t=1}^{T} \left\langle \pi^*(\cdot|x) - \pi_t(\cdot|x), q_t(x, \cdot) \right\rangle \leq \frac{\mathcal{D}\left(\pi^*(\cdot|x) \| \pi_1(\cdot|x)\right)}{\alpha} + \frac{\alpha \left\| q_t(x, \cdot) \right\|_\infty^2}{2}.$$

Finally, using that $\|q_t\|_\infty \leq D_q$ and taking an expectation with respect to $x \sim \nu^{\pi^*}$ concludes the proof. □

E DETAILED COMPUTATIONS FOR COMPARING COVERAGE RATIOS

In this section, after reviewing the different versions of coverage ratio discussed in the paper, we prove several inequalities that hold between them. For ease of comparison, we only consider discounted linear MDPs (Definition 2.1).

Definition E.1. Recall the following definitions of coverage ratio given by different authors in the offline RL literature:

1. $C_{\varphi,c}(\pi^*;\pi_B) = \mathbb{E}_{X,A\sim\mu^*}\left[\boldsymbol{\varphi}(X,A)\right]^\top \boldsymbol{\Lambda}^{-2c}\mathbb{E}_{X,A\sim\mu^*}\left[\boldsymbol{\varphi}(X,A)\right]$ (Ours)
2. $C^\circ(\pi^*;\pi_B) = \mathbb{E}_{X,A\sim\mu^*}\left[\sqrt{\boldsymbol{\varphi}(X,A)^\intercal\boldsymbol{\Lambda}^{-1}\boldsymbol{\varphi}(X,A)}\right]$ (e.g., Jin et al. (2021))
3. $C^\diamond(\pi^*;\pi_B) = \mathbb{E}_{X,A\sim\mu^*}\left[\boldsymbol{\varphi}(X,A)^\intercal\boldsymbol{\Lambda}^{-1}\boldsymbol{\varphi}(X,A)\right]$ (e.g., Gabbianelli et al. (2023))
4. $C^\dagger(\pi^*;\pi_B) = \sup_{y\in\mathbb{R}^d}\frac{y^\intercal\mathbb{E}_{X,A\sim\mu^*}\left[\boldsymbol{\varphi}(X,A)\boldsymbol{\varphi}(X,A)^\intercal\right]y}{y^\intercal\mathbb{E}_{X,A\sim\mu_B}[\boldsymbol{\varphi}(X,A)\boldsymbol{\varphi}(X,A)^\intercal]y}$ (e.g., Uehara and Sun (2022))
5. $C_{\mathcal{F},\pi}(\pi^*;\pi_B) = \max_{f\in\mathcal{F}}\frac{\|f-\mathcal{T}^\pi f\|^2_{\mu^*}}{\|f-\mathcal{T}^\pi f\|^2_{\mu_B}}$ (e.g., Xie et al. (2021)),

where $c\in\{1,2\}$, $\boldsymbol{\Lambda} = \mathbb{E}_{X,A\sim\mu_B}\left[\boldsymbol{\varphi}(X,A)\boldsymbol{\varphi}(X,A)^\intercal\right]$ (assumed invertible), $\mathcal{F}\subseteq\mathbb{R}^{\mathcal{X}\times\mathcal{A}}$, and $\mathcal{T}^\pi:\mathcal{F}\to\mathbb{R}$ defined as $(\mathcal{T}^\pi f)(x,a) = r(x,a) + \gamma\sum_{x',a'}p(x'|x,a)\pi(a'|x')f(x',a')$ is the Bellman operator associated to policy π.

Remark E.2. By Jensen's inequality, it is clear that $C^\circ \le \sqrt{C^\diamond}$. However, $C^\diamond$ is conceptually similar to the more common C° and allows for interesting comparisons with the other notions of coverage, as shown later in this section.

In the following, we construct a problem instance where C° can be arbitrarily larger than $C_{\varphi,c}$, regardless of the value of c, thanks to the single-direction property of our coverage ratio discussed in Section 6.

Proposition E.3. *There exists a linear MDP with two states, two actions and feature dimension $d=3$, such that, for every $\epsilon\in(0,1)$, there exists a behavior policy π_B, such that $C_{\varphi,c}(\pi^*;\pi_B)$ is bounded by a constant independent of ϵ for all $c\in\{1/2,1\}$, while $C^\circ(\pi^*;\pi_B) = \Omega(\epsilon^{-1/2})$, where π^* is the unique deterministic optimal policy of the MDP.*

Proof. Let $|\mathcal{X}| = \{x_1,x_2\}$ and $\mathcal{A} = \{a_1,a_2\}$. Consider the following 3-dimensional feature map where $\boldsymbol{\varphi}_{ij}$ is short for $\boldsymbol{\varphi}(x_i,a_j)$:

$$\boldsymbol{\varphi}_{11} = [4,0,1]^\top, \qquad \boldsymbol{\varphi}_{12} = [1,1,1]^\top,$$
$$\boldsymbol{\varphi}_{21} = [0,4,1]^\top, \qquad \boldsymbol{\varphi}_{22} = [-1,-1,1]^\top.$$

Following the notation of Definition 2.1, let $\boldsymbol{\psi}(x_1) = \boldsymbol{\psi}(x_2) = [0,0,1/2]^\top$ and $\boldsymbol{\omega} = [1,1,0]^\top$, obtaining $p(x_k|x_i,a_j) = 1/2$ for all $i,j,k\in[2]$, and the following reward function:

$$r(x_1,a_1) = 4, \qquad r(x_1,a_2) = 2,$$
$$r(x_2,a_1) = 4, \qquad r(x_2,a_2) = -2.$$

Finally, let $\nu_0(x_1) = \nu_0(x_2) = 1/2$. It is easy to see that, for any discount factor $\gamma>0$, the MDP admits a unique deterministic optimal policy, $\pi^*(x_1) = \pi^*(x_2) = a_1$, with optimal value $\rho^* = 4(1-\gamma)$. The state-action occupancy measure induced by this optimal policy is

$$\mu^*(x_1,a_1) = \mu^*(x_2,a_1) = \frac{1}{2}, \qquad \mu^*(x_1,a_2) = \mu^*(x_2,a_2) = 0.$$

Now fix an $\epsilon\in(0,1)$. Let the behavior policy be

$$\pi_B(a_1|x_1) = \epsilon, \qquad \pi_B(a_2|x_1) = 1-\epsilon,$$
$$\pi_B(a_1|x_2) = \epsilon, \qquad \pi_B(a_2|x_2) = 1-\epsilon.$$

The state-action occupancy measure induced by the behavior policy is

$$\mu_B(x_1, a_1) = \frac{\epsilon}{2}, \qquad \mu_B(x_1, a_2) = \frac{1-\epsilon}{2},$$
$$\mu_B(x_2, a_1) = \frac{\epsilon}{2}, \qquad \mu_B(x_2, a_2) = \frac{1-\epsilon}{2}.$$

The feature covariance matrix under π_b is then

$$\mathbf{\Lambda} = \mathbb{E}_{X,A\sim\mu_B}[\boldsymbol{\varphi}(X,A)\boldsymbol{\varphi}(X,A)^\top] = \begin{bmatrix} 1+7\epsilon & 1-\epsilon & 2\epsilon \\ 1-\epsilon & 1+7\epsilon & 2\epsilon \\ 2\epsilon & 2\epsilon & 1 \end{bmatrix},$$

from which we obtain the coverage ratio

$$C^\circ(\pi^*; \pi_B) = \mathbb{E}_{X,A\sim\mu^*}\left[\sqrt{\boldsymbol{\varphi}(X,A)^\intercal \mathbf{\Lambda}^{-1}\boldsymbol{\varphi}(X,A)}\right] = \sqrt{\frac{1+9\epsilon}{\epsilon(1+4\epsilon)}} = \Omega(\epsilon^{-1/2}). \tag{29}$$

To compute $C_{\varphi,c}(\pi^*; \pi_B)$, note that the expected feature vector under π^* is

$$\overline{\boldsymbol{\varphi}}(\pi^*) = \mathbb{E}_{X,A\sim\mu^*}\left[\boldsymbol{\varphi}(X,A)\right] = [2,2,1]^\top.$$

Hence:

$$C_{\varphi,1/2}(\pi^*; \pi_B) = \overline{\boldsymbol{\varphi}}(\pi^*)^\top \mathbf{\Lambda}^{-1}\overline{\boldsymbol{\varphi}}(\pi^*) = \frac{5}{1+4\epsilon} \leq 5, \tag{30}$$

$$C_{\varphi,1}(\pi^*; \pi_B) = \overline{\boldsymbol{\varphi}}(\pi^*)^\top \mathbf{\Lambda}^{-2}\overline{\boldsymbol{\varphi}}(\pi^*) = \frac{3}{(1+4\epsilon)^2} \leq 3 < 5. \tag{31}$$

□

The previous proof admits a simple geometric interpretation: for $\epsilon \to 0$, the span of the features visited by the behavior policy degenerates to $\text{span}(\{\boldsymbol{\varphi}_{12}, \boldsymbol{\varphi}_{22}\})$, which belongs to a 2-dimensional subspace of $\mathbb{R}^3$, while the optimal features span the whole $\mathbb{R}^3$. So, according to the notion of coverage from Jin et al. (2021), the data fail to cover the span of the optimal features. However, the average optimal feature $\overline{\boldsymbol{\varphi}}(\pi^*)$ belongs to the very same subspace covered by the data, which is enough according to our notion of coverage. In particular, $\overline{\boldsymbol{\varphi}}(\pi^*) = 3/2\boldsymbol{\varphi}_{12} - 1/2\boldsymbol{\varphi}_{22}$.

The following is a generalization of the low-variance property discussed in Section 6.

Proposition E.4. *Let $\mathbb{V}[Z] = \mathbb{E}[\|Z - \mathbb{E}[Z]\|^2]$ for a random vector Z. Then, for any pair of policies π^*, π_B*

$$C_{\varphi,c}(\pi^*; \pi_B) = \mathbb{E}_{X,A\sim\mu^*}\left[\boldsymbol{\varphi}(X,A)^\intercal \mathbf{\Lambda}^{-2c}\boldsymbol{\varphi}(X,A)\right] - \mathbb{V}_{X,A\sim\mu^*}\left[\mathbf{\Lambda}^{-c}\boldsymbol{\varphi}(X,A)\right].$$

In particular, $C_{\varphi,1/2}(\pi^*; \pi_B) \leq C^\diamond(\pi^*; \pi_B)$ for all π^*, π_B.

Proof. We just rewrite $C_{\varphi,c}$ from Definition E.1 as

$$C_{\varphi,c}(\pi^*; \pi_B) = \left\|\mathbb{E}_{X,A\sim\mu^*}\left[\mathbf{\Lambda}^{-c}\boldsymbol{\varphi}(X,A)\right]\right\|^2.$$

The result follows from the elementary property of variance $\mathbb{V}[Z] = \mathbb{E}[\|Z\|^2] - \|\mathbb{E}[Z]\|^2$. The second statement follows from the non-negativity of the variance, but can also be obtained directly via Jensen's inequality. □

Proposition E.5. $C^\dagger(\pi^*; \pi_B) \leq C^\diamond(\pi^*; \pi_B) \leq dC^\dagger(\pi^*; \pi_B)$.

Proof. Let $(X^*, A^*) \sim \mu^*$ and $\boldsymbol{M} = \mathbb{E}[\boldsymbol{\varphi}(X^*, A^*)\boldsymbol{\varphi}(X^*, A^*)]$. First, we rewrite $C^\diamond$ as

$$\begin{aligned} C^\diamond(\pi^*; \pi_B) &= \mathbb{E}\left[\boldsymbol{\varphi}(X^*, A^*)^\intercal \mathbf{\Lambda}^{-1}\boldsymbol{\varphi}(X^*, A^*)\right] \\ &= \mathbb{E}\left[\text{Tr}(\boldsymbol{\varphi}(X^*, A^*)^\intercal \mathbf{\Lambda}^{-1}\boldsymbol{\varphi}(X^*, A^*))\right] \\ &= \mathbb{E}\left[\text{Tr}(\boldsymbol{\varphi}(X^*, A^*)\boldsymbol{\varphi}(X^*, A^*)^\intercal \mathbf{\Lambda}^{-1})\right] \qquad (32) \\ &= \text{Tr}(\boldsymbol{M}\mathbf{\Lambda}^{-1}) \qquad (33) \\ &= \text{Tr}(\mathbf{\Lambda}^{-1/2}\boldsymbol{M}\mathbf{\Lambda}^{-1/2}), \qquad (34) \end{aligned}$$

where we have used the cyclic property of the trace (twice) and linearity of trace and expectation. Note that, since $\mathbf{\Lambda}$ is positive definite, it admits a unique positive definite matrix $\mathbf{\Lambda}^{1/2}$ such that $\mathbf{\Lambda} = \mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}$. We rewrite $C^\dagger$ in a similar fashion

$$\begin{aligned} C^\dagger(\pi^*; \pi_B) &= \sup_{y \in \mathbb{R}^d} \frac{y^\intercal \boldsymbol{M} y}{y^\intercal \mathbf{\Lambda} y} \\ &= \sup_{z \in \mathbb{R}^d} \frac{z^\intercal \mathbf{\Lambda}^{-1/2} \boldsymbol{M} \mathbf{\Lambda}^{-1/2} z}{z^\intercal z} \qquad (35) \\ &= \lambda_{\max}(\mathbf{\Lambda}^{-1/2} \boldsymbol{M} \mathbf{\Lambda}^{-1/2}), \qquad (36) \end{aligned}$$

where $\lambda_{\max}$ denotes the maximum eigenvalue of a matrix. We have used the fact that both $\boldsymbol{M}$ and $\mathbf{\Lambda}$ are positive definite and the min-max theorem. Since the quadratic form $\mathbf{\Lambda}^{-1/2} \boldsymbol{M} \mathbf{\Lambda}^{-1/2}$ is also positive definite, and the trace is the sum of the (positive) eigenvalues, we get the desired result. □

Proposition E.6 (cf. the proof of Theorem 3.2 from (Xie et al., 2021))**.** *Let $\mathcal{F} = \{f_{\boldsymbol{\theta}} : (x, a) \mapsto \langle \boldsymbol{\varphi}(x, a), \boldsymbol{\theta} \rangle | \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\}$ where φ is the feature map of the linear MDP. Then*

$$C_{\mathcal{F},\pi}(\pi^*; \pi_B) \leq C^\dagger(\pi^*; \pi_B),$$

with equality if $\Theta = \mathbb{R}^d$.

Proof. Fix any policy π and let $\mathcal{T} = \mathcal{T}^\pi$. By linear Bellman completeness of linear MDPs (Jin et al., 2020), $\mathcal{T} f \in \mathcal{F}$ for any $f \in \mathcal{F}$. For $f_{\boldsymbol{\theta}} : (x, a) \mapsto \langle \boldsymbol{\varphi}(x, a), \boldsymbol{\theta} \rangle$, let $\mathcal{T}\boldsymbol{\theta} \in \Theta$ be defined so that $\mathcal{T} f_{\boldsymbol{\theta}} : (x, a) \mapsto \langle \boldsymbol{\varphi}(x, a), \mathcal{T}\boldsymbol{\theta} \rangle$. Then

$$\begin{aligned} C_{\mathcal{F},\pi}(\pi^*; \pi_B) &= \max_{f \in \mathcal{F}} \frac{\mathbb{E}_{X,A \sim \mu^*}\left[(f(X, A) - \mathcal{T} f(X, A))^2\right]}{\mathbb{E}_{X,A \sim \mu_B}\left[(f(X, A) - \mathcal{T} f(X, A))^2\right]} \qquad (37) \\ &\leq \max_{\boldsymbol{\theta} \in \mathbb{R}^d} \frac{\mathbb{E}_{X,A \sim \mu^*}\left[\langle \boldsymbol{\varphi}(X, A), \boldsymbol{\theta} - \mathcal{T}\boldsymbol{\theta} \rangle^2\right]}{\mathbb{E}_{X,A \sim \mu_B}\left[\langle \boldsymbol{\varphi}(X, A), \boldsymbol{\theta} - \mathcal{T}\boldsymbol{\theta} \rangle^2\right]} \qquad (38) \\ &= \max_{y \in \mathbb{R}^d} \frac{\mathbb{E}_{X,A \sim \mu^*}\left[\langle \boldsymbol{\varphi}(X, A), y \rangle^2\right]}{\mathbb{E}_{X,A \sim \mu_B}\left[\langle \boldsymbol{\varphi}(X, A), y \rangle^2\right]} \qquad (39) \\ &= \max_{y \in \mathbb{R}^d} \frac{y^\intercal \mathbb{E}_{X,A \sim \mu^*}\left[\boldsymbol{\varphi}(X, A) \boldsymbol{\varphi}(X, A)^\intercal\right] y}{y^\intercal \mathbb{E}_{X,A \sim \mu_B}\left[\boldsymbol{\varphi}(X, A) \boldsymbol{\varphi}(X, A)^\intercal\right] y}, \qquad (40) \end{aligned}$$

where the inequality in Equation (38) holds with equality if $\Theta = \mathbb{R}^d$. □