Efficient Global Planning in Large MDPs
via Stochastic Primal-Dual Optimization

Gergely Neu   GERGELY.NEU@GMAIL.COM
Universitat Pompeu Fabra, Barcelona, Spain

Nneka Okolo   NNEKAMAUREEN.OKOLO@UPF.EDU
Universitat Pompeu Fabra, Barcelona, Spain

Editors: Shipra Agrawal and Francesco Orabona

Abstract

We propose a new stochastic primal-dual optimization algorithm for planning in a large discounted Markov decision process with a generative model and linear function approximation. Assuming that the feature map approximately satisfies standard realizability and Bellman-closedness conditions and also that the feature vectors of all state-action pairs are representable as convex combinations of a small core set of state-action pairs, we show that our method outputs a near-optimal policy after a polynomial number of queries to the generative model. Our method is computationally efficient and comes with the major advantage that it outputs a single softmax policy that is compactly represented by a low-dimensional parameter vector, and does not need to execute computationally expensive local planning subroutines in runtime.

Keywords: Markov decision processes, Linear Programming, Linear function approximation, Planning with a generative model.

1. Introduction

Finding near-optimal policies in large Markov decision processes (MDPs) is one of the most important tasks encountered in model-based reinforcement learning Sutton and Barto (2018). This problem (commonly referred to as *planning*) presents both computational and statistical challenges when having access only to a generative model of the environment: an efficient planning method should use little computation and few samples drawn from the model. While the complexity of planning in small Markov decision processes is already well-understood by now (cf. Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2020b), extending the insights to large state spaces with function approximation has remained challenging. One particular challenge that remains unaddressed in the present literature is going beyond *local planning* methods that require costly online computations for producing each action during execution time. In this work, we advance the state of the art in planning for large MDPs by designing and analyzing a new *global planning* algorithm that outputs a simple, compactly represented policy after performing all of its computations in an offline fashion.

Our method is rooted in the classic linear programming framework for sequential decision making first proposed by Manne (1960); d'Epenoux (1963); Denardo (1970). This approach formulates the problem of finding an optimal behavior policy as a linear program (LP) with a number of variables and constraints that is proportional to the size of the state-action space of the MDP. Over the last several decades, numerous attempts have been made to derive computationally tractable algorithms from the LP framework, most notably via reducing the number of variables using function approximation techniques. This idea has been first explored by Schweitzer and Seidmann (1985),

whose ideas were further developed and popularized by the seminal work of De Farias and Van Roy (2003). Further approximations were later proposed by De Farias and Van Roy (2004); Petrik and Zilberstein (2009); Lakshminarayanan et al. (2017), whose main focus was reducing the complexity of the LPs while keeping the quality of optimal solutions under control.

More recently, the LP framework has started to inspire practical methods for reinforcement learning and, more specifically, planning with a generative model. The most relevant to the present paper is the line of work initiated by Wang and Chen (2016); Wang (2017), who proposed planning algorithms based on primal-dual optimization of the Lagrangian associated with the classic LP. Over the last several years, this computational recipe has been perfected for small MDPs by works like Cheng et al. (2020); Jin and Sidford (2020); Tiapkin and Gasnikov (2022), and even some extensions using linear function approximation have been proposed by Chen et al. (2018); Bas-Serrano and Neu (2020). While these methods have successfully reduced the number of primal and dual variables, they all require stringent conditions on the function approximator being used and their overall sample complexity scales poorly with the size of the MDP. For instance, the bounds of Chen et al. (2018) scale with a so-called "concentrability coefficient" that can be as large as the size of the state space, thus failing to yield meaningful guarantees for large MDPs. Furthermore, these methods require parametrizing the space of so-called state-action occupancy measures, which is undesirable given the complexity of said space.

In the present work, we develop a new stochastic primal-dual method based on a relaxed version of the classical LP. This relaxed LP (inspired by the works of Mehta and Meyn, 2009; Neu and Pike-Burke, 2020 and Bas-Serrano et al., 2021) features a linearly parametrized action-value function and guarantees to produce optimal policies as solutions under well-studied conditions like the linear MDP assumption of Yang and Wang (2019); Jin et al. (2020). Our method iteratively updates the primal and dual variables via a combination of stochastic mirror descent steps and a set of implicit update rules tailor-made for the relaxed LP. Under a so-called *core state-action-pair* assumption, we show that the method produces a near-optimal policy with sample and computation complexity that is polynomial in the relevant problem parameters: the size of the core set, the dimensionality of the feature map, the effective horizon, and the desired accuracy. Additional assumptions required by our analysis are the near-realizability of the Q-functions and closedness under the Bellman operators of all policies. The main merit of our algorithm is that it produces a compactly represented softmax policy which requires no access to the generative model in runtime.

The works most directly comparable to ours are Wang et al. (2021), Shariff and Szepesvári (2020), and Yin et al. (2022)—their common feature being the use of a core set of states or state-action pairs for planning. Wang et al. (2021) provide a sample complexity bound for finding a near-optimal policy in linear MDPs, but their algorithm is computationally intractable[1] due to featuring a planning subroutine in a linear MDP. Shariff and Szepesvári (2020) and Yin et al. (2022) both propose local planning methods that require extensive computation in runtime, but on the other hand some of their assumptions are weaker than ours. In particular, they only require realizability of the value functions but continue to operate without the function class being closed under all Bellman operators. For a detailed discussion on the role and necessity of such assumptions, we refer to the excellent discussion provided by Yin et al. (2022) on the subject.

---

1. At least we are not aware of a computationally efficient global planning method that works for the discounted case they consider.

Notation. In the n-dimensional Euclidean space $\mathbb{R}^n$, we denote the vector of all ones by $\mathbf{1}$, the zero vector by $\mathbf{0}$ and the i-th coordinate vector by $\boldsymbol{e_i}$. For, vectors $a, b \in \mathbb{R}^m$, we use $a \leq b$ to denote elementwise comparison, meaning that $a_i \leq b_i$ is satisfied for all entries i. For any finite set D, $\Delta_D = \{p \in \mathbb{R}^{|D|}_+ \mid \|p\|_1 = 1\}$ denotes the set of all probability distributions over its entries. We define the relative entropy between two distributions $p, p' \in \Delta_D$ as $\mathcal{D}(p\|p') = \sum_{i=1}^{|D|} p_i \log \frac{p_i}{p'_i}$. In the context of iterative algorithms, we will use the notation $\mathcal{F}_{t-1}$ to refer to the sigma-algebra generated by all events up to the end of iteration $t - 1$, and use the shorthand notation $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ to denote expectation conditional on the past history encountered by the method.

2. Preliminaries

We study a discounted Markov Decision Processes (Puterman, 2014) denoted by the quintuple $(\mathcal{X}, \mathcal{A}, r, P, \gamma)$ with $\mathcal{X}$ and $\mathcal{A}$ representing finite (but potentially very large) state and action spaces of cardinality $X = |\mathcal{X}|, A = |\mathcal{A}|$ respectively. The reward function is denoted by $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, and the transition function by $P : \mathcal{X} \times \mathcal{A} \to \Delta_{\mathcal{X}}$. We will often represent the reward function by a vector in $\mathbb{R}^{XA}$ and the transition function by the operator $P \in \mathbb{R}^{XA \times X}$ which acts on functions $v \in \mathbb{R}^X$ by assigning $(Pv)(x, a) = \sum_{x'} P(x'|x, a)v(x')$ for all x, a. Its adjoint $P^\mathsf{T} \in \mathbb{R}^{X \times XA}$ is similarly defined on functions $u \in \mathbb{R}^{XA}$ via the assignment $(P^\mathsf{T} u)(x) = \sum_{x', a'} P(x|x', a')u(x', a')$ for all x. We also define the operator $E \in \mathbb{R}^{XA \times X}$ and its adjoint $E^\mathsf{T} \in \mathbb{R}^{X \times XA}$ acting on respective vectors $v \in \mathbb{R}^X$ and $u \in \mathbb{R}^{XA}$ through the assignments $(Ev)(x, a) = v(x)$ and $(E^\mathsf{T} u)(x) = \sum_a u(x, a)$. For simplicity, we assume the rewards are bounded in $[0, 1]$ and let $\mathcal{Z} = \{(x, a) | x \in \mathcal{X}, a \in \mathcal{A}\}$ denote the set of all possible state action pairs with cardinality $Z = |\mathcal{Z}|$ to be used when necessary.

The Markov decision process describes a sequential decision-making process where in each round $t = 0, 1, \ldots$, the *agent* observes the state of the environment x_t, takes an action a_t, and earns a potentially random reward with expectation $r(x_t, a_t)$. The state of the environment in the next round $t + 1$ is generated randomly according to the transition dynamics as $x_{t+1} \sim P(\cdot|x_t, a_t)$. The initial state of the process x_0 is drawn from a fixed distribution $\nu_0 \in \Delta_{\mathcal{X}}$. In the setting we consider, the agent's goal is to maximize its normalized discounted return $(1-\gamma)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t)\right]$, where $\gamma \in (0, 1)$ is the discount factor, and the expectation is taken over the random transitions generated by the environment, the random initial state, and the potential randomness injected by the agent. It is well-known that maximizing the discounted return can be achieved by following a memoryless time-independent decision making rule mapping states to actions. Thus, we restrict our attention to *stationary stochastic policies* $\pi : \mathcal{X} \to \Delta_{\mathcal{A}}$ with $\pi(a|x)$ denoting the probability of the policy taking action a in state x. We define the mean operator $M^\pi : \mathbb{R}^{XA} \to \mathbb{R}^X$ with respect to policy π that acts on functions $Q \in \mathbb{R}^{X \times A}$ as $(M^\pi Q)(x) = \sum_a \pi(a|x)Q(x, a)$ and the (non-linear) max operator $M : \mathbb{R}^{XA} \to \mathbb{R}^X$ that acts as $(M^* Q)(x) = \max_{a \in \mathcal{A}} Q(x, a)$. The value function and action-value function of policy π are respectively defined as $V^\pi(x) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) \big| x_0 = x\right]$ and $Q^\pi(x, a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) \big| x_0 = x, a_0 = a\right]$, where the notation $\mathbb{E}_\pi[\cdot]$ signifies that each action is selected by following the policy as $a_t \sim \pi(\cdot|x_t)$. The value functions of a policy π are known to satisfy the Bellman equations (Bellman, 1966) and the value functions of an optimal policy π^* satisfy the Bellman optimality equations, which can be conveniently written in our notation as

$$Q^\pi = r + \gamma P V^\pi \qquad \text{and} \qquad Q^* = r + \gamma P V^*, \qquad (1)$$

with $V^\pi = M^\pi Q^\pi$ and $V^* = M^* Q^*$. Any optimal policy satisfies $\text{supp}(\pi(\cdot|x)) \subseteq \arg\max_a Q^*(x,a)$. We refer the reader to Puterman (2014) for the standard proofs of these fundamental results.

Our approach taken in this paper will primarily make use of an alternative formulation of the MDP optimization based on linear programming, due to Manne (1960) (see also d'Epenoux, 1963, Denardo, 1970, and Section 6.9 in Puterman, 2014). This formulation phrases the optimal control problem as a search for an *occupancy measure* with maximal return. The state-action occupancy measure of a policy π is defined as

$$\mu^\pi(x,a) = (1-\gamma)\mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t \mathbb{I}_{\{(x_t,a_t)=(x,a)\}}\right],$$

which allows rewriting the expected normalized return of π as $R_\gamma^\pi = \langle \mu^\pi, r \rangle$. The corresponding state-occupancy measure is defined as the state distribution $\nu^\pi(x) = \sum_a \mu^\pi(x,a)$. Denoting the direct product of a state distribution ν and the policy π by $\nu \circ \pi$ with its entries being $(\nu \circ \pi)(x,a) = \nu(x)\pi(a|x)$, we can notice that the state-action occupancy measure induced by π satisfies $\mu^\pi = \nu^\pi \circ \pi$. The set of all occupancy measures can be fully characterized by a set of linear constraints, which allows rewriting the optimal control problem as the following linear program (LP):

$$\begin{aligned}
\max_{\mu \in \mathbb{R}_+^{XA}} \quad & \langle \mu, r \rangle \\
\text{subject to} \quad & E^\mathsf{T}\mu = (1-\gamma)\nu_0 + \gamma P^\mathsf{T}\mu.
\end{aligned} \tag{2}$$

Any optimal solution μ^* of this LP can be shown to correspond to the occupancy measure of an optimal policy π^*, and generally policies can be extracted from feasible points μ via the rule $\pi_\mu(a|x) = \mu(x,a)/\sum_{a'} \mu(x,a')$ (subject to the denominator being nonzero). The dual of this linear program takes the following form:

$$\begin{aligned}
\min_{V \in \mathbb{R}^X} \quad & (1-\gamma)\langle \nu_0, V \rangle \\
\text{subject to} \quad & EV \geq r + \gamma PV.
\end{aligned} \tag{3}$$

The optimal value function V^* is known to be an optimal solution for this LP, and is the unique optimal solution provided that ν_0 has full support over the state space.

The above linear programs are not directly suitable as solution tools for large MDPs, given that they have a large number of variables and constraints. Numerous adjustments have been proposed over the last decades to address this issue (Schweitzer and Seidmann, 1985; De Farias and Van Roy, 2003, 2004; Lakshminarayanan et al., 2017; Bas-Serrano and Neu, 2020). One recurring theme in these alternative formulations is relaxing some of the constraints by the introduction of a *feature map*. In this work, we use as starting point an alternative proposed by Bas-Serrano et al. (2021), who use a feature map $\varphi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ represented by the $XA \times d$ *feature matrix* Φ, and rephrase the original optimization problem as the following LP:

$$\begin{aligned}
\max_{\mu,u \in \mathbb{R}_+^{XA}} \quad & \langle \mu, r \rangle \\
\text{subject to} \quad & E^\mathsf{T}u = (1-\gamma)\nu_0 + \gamma P^\mathsf{T}\mu, \\
& \Phi^\mathsf{T}\mu = \Phi^\mathsf{T}u \,.
\end{aligned} \tag{4}$$

The dual of this LP is given as

$$\min_{\theta \in \mathbb{R}^d, V \in \mathbb{R}^X} \ (1 - \gamma)\langle \nu_0 , V \rangle$$
$$\text{subject to} \quad \Phi\theta \geq r + \gamma P V,$$
$$E V \geq \Phi\theta \, .$$

As shown by Bas-Serrano et al. (2021), optimal solutions of the above relaxed LPs correspond to optimal occupancy measures and value functions under the popular linear MDP assumption of Yang and Wang (2019); Jin et al. (2020). The key merit of these LPs is that the dual features a linearly parametrized action-value function $Q_\theta = \Phi\theta$, which allows the extraction of a simple greedy policy $\pi_\theta(a|x) = \mathbb{I}_{\{a = \arg\max_{a'} Q_\theta(x, a')\}}$ from any dual solution θ, and in particular an optimal policy can be extracted from its optimal solution.

3. A tractable linear program for large MDPs with linear function approximation

The downside of the classical linear programs defined in the previous sections is that they all feature a large number of variables and constraints, even after successive relaxations. In what follows, we offer a new version of the LP (4) that gets around this issue and allows the development of a tractable primal-dual algorithm with strong performance guarantees. In particular, we will reduce the number of variables in the primal LP (4) by considering only sparsely supported state-action distributions instead of full occupancy measures, which will be justified by the following technical assumption made on the MDP structure:

Assumption 1 *(Approximate Core State-Action Assumption) The feature vector of any state-action pair $(x, a) \in \mathcal{Z}$ can be approximately expressed as a convex combination of features evaluated at a set of m core state-action pairs $(x', a') \in \widetilde{\mathcal{Z}}$, up to an error $\Delta_{\text{core}} \in \mathbb{R}^{XA \times d}$. That is, for each $(x, a) \in \mathcal{Z}$, there exists a set of coefficients satisfying $b(x', a'|x, a) \geq 0$ and $\sum_{x', a'} b(x', a'|x, a) = 1$ such that $\varphi(x, a) = \sum_{x', a'} b(x', a'|x, a)\varphi(x', a') + \Delta_{\text{core}}(x, a)$. Furthermore, for every x, a, the misspecification error satisfies $\|\Delta_{\text{core}}(x, a)\|_2 = \varepsilon_{\text{core}}(x, a)$.*

It will be useful to rephrase this assumption using the following handy notation. Let $\mathcal{U} \in \mathbb{R}_+^{m \times Z}$ denote a selection matrix such that, $\widetilde{\Phi} = \mathcal{U}\Phi \in \mathbb{R}^{m \times d}$ is the core feature matrix with rows corresponding to Φ evaluated at core state-action pairs. Furthermore, the interpolation coefficients from Assumption 1 can be organized into a stochastic matrix $\mathcal{B} \in \mathbb{R}^{Z \times m}$ with $\mathcal{B}(x, a) = \{b(x', a'|x, a)\}_{(x', a') \in \widetilde{\mathcal{Z}}} \in \mathbb{R}_+^m$ for $(x, a) \in \mathcal{Z}$. Then, the assumption can be simply rephrased as requiring the condition that $\Phi = \mathcal{B}\mathcal{U}\Phi + \Delta_{\text{core}}$. Note that both $\mathcal{U}$ and $\mathcal{B}$ are stochastic matrices satisfying $\mathcal{U}\mathbf{1} = \mathbf{1}$ and $\mathcal{B}\mathbf{1} = \mathbf{1}$, and the same holds for their product $\mathcal{B}\mathcal{U}\mathbf{1} = \mathbf{1}$. Note however that $\mathcal{B}\mathcal{U}$ is a rank-m matrix, which implies that the assumption can only be satisfied with zero error whenever $m \geq \text{rank}(\Phi)$, which in general can be as large as the feature dimensionality d. Whether or not it is possible to find a set of core-state-action pairs in a given MDP is a nontrivial question that we discuss in Section 6.

With the notation introduced above, we are ready to state our relaxation of the LP (4) that serves as the foundation for our algorithm design:

$$\max_{\lambda \in \mathbb{R}_+^m, u \in \mathbb{R}_+^{XA}} \ \langle \lambda , \mathcal{U} r \rangle$$
$$\text{subject to} \quad E^{\mathsf{T}} u = (1 - \gamma)\nu_0 + \gamma P^{\mathsf{T}} \mathcal{U}^{\mathsf{T}} \lambda, \tag{5}$$
$$\Phi^{\mathsf{T}} \mathcal{U}^{\mathsf{T}} \lambda = \Phi^{\mathsf{T}} u \, .$$

The dual of the LP can be written as follows:

$$\min_{\theta \in \mathbb{R}^d, V \in \mathbb{R}^X} (1 - \gamma)\langle \nu_0, V \rangle$$
$$\text{subject to } EV \geq Q_\theta,$$
$$\mathcal{U}Q_\theta \geq \mathcal{U}(r + \gamma PV). \tag{6}$$

The above LPs can be shown to yield optimal solutions that correspond to optimal occupancy measures and value functions under a variety of conditions. The first of these is the so-called linear MDP condition due to Yang and Wang (2019); Jin et al. (2020), which is recalled below:

Definition 1 *(Linear MDP) An MDP is called a linear MDP if there exists $W \in \mathbb{R}^{d \times X}$ and $\vartheta \in \mathbb{R}^d$ such that, the transition matrix P and reward vector r can be written as the linear functions $P = \Phi W$ and $r = \Phi \vartheta$.*

It is easy to see that the relaxed LPs above retain the optimal solutions of the original LP as long as $\Delta_{\text{core}} = 0$—we provide the straightforward proof in Appendix A.1. As can be seen from the Bellman equations (1), the action-value function of any policy π can be written as $Q^\pi = \Phi \theta^\pi$ for some θ^π under the linear MDP assumption. The linearity of the transition function is a very strong condition that is often not satisfied in problems of practical interest, which motivates us to study a more general class of MDPs with weaker feature maps. The following two concepts will be useful in characterizing the power of the feature map.

Roughly speaking, the conditions below correspond to supposing that all Q-functions can be *approximately* represented by some parameter vectors with bounded norms, and similarly the Bellman operators applied to feasible Q-functions are also approximately representable with linear functions with bounded coefficients. Concretely, we will assume that all feature vectors satisfy $\|\varphi(x, a)\|_2 \leq R$ for all (x, a) and will work with parameter vectors with norm bounded as $\|\theta\|_2 \leq D_\gamma$, whose set will be denoted as $\mathbb{B}(D_\gamma) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq D_\gamma\}$. The first property of interest we define here characterizes the best approximation error that this set of parameters and features has in terms of representing the action-value functions:

Definition 2 *(Q-Approximation Error) The Q-approximation error associated with a policy π is defined as $\epsilon_\pi = \inf_{\theta \in \mathbb{B}(D_\gamma)} \|Q^\pi - \Phi \theta\|_\infty$.*

Since the true action-value functions of any policy satisfy $\|Q^\pi\|_\infty \leq \frac{1}{1-\gamma}$ and members of our function class satisfy $\|Q_\theta\|_\infty \leq RD_\gamma$, the bound D_γ should scale linearly with $\frac{1}{1-\gamma}$ to accommodate all potential Q-functions with small error. The second property of interest is the ability of the feature map to capture applications of the Bellman operator to functions within the approximating class:

Definition 3 *(Inherent Bellman Error) The* Bellman Error *(BE) associated with the pair of parameter vectors $\theta, \theta' \in \mathbb{B}(D_\gamma)$ and a policy π is defined as the state-action vector $\text{BE}^\pi(\theta, \theta') \in \mathbb{R}^{XA}$ with components*

$$\text{BE}^\pi(\theta, \theta') = r + \gamma P M^\pi Q_{\theta'} - Q_\theta.$$

The Inherent Bellman Error *is then defined as* $\text{IBE} = \sup_\pi \sup_{\theta' \in \mathbb{B}(D_\gamma)} \inf_{\theta \in \mathbb{B}(D_\gamma)} \|\text{BE}^\pi(\theta, \theta')\|_\infty$.

Function approximators with zero IBE are often called "Bellman complete" and have been intensely studied in reinforcement learning (Antos et al., 2008; Chen and Jiang, 2019; Zanette et al., 2020)— note however that our notion is stronger than some others appearing in these works as it requires

small error for all policies π and not just greedy ones. As shown by Jin et al. (2020), linear MDP models enjoy zero inherent Bellman error in the stricter sense used in the above definition. They also establish that the converse is also true: having zero IBE for all policies implies linearity of the transition model. Provided that D_γ is set large enough, one can show that feature maps with zero IBE satisfy $\varepsilon_\pi = 0$ for all policies π, and optimal solutions of the above relaxed LPs yield optimal policies when $\Delta_{\mathrm{core}} = 0$. The interested reader can verify this by appropriately adjusting the arguments in Appendix A.1 (see Footnote 3).

4. Algorithm and main results

We now turn to presenting our main contribution: a computationally efficient planning algorithm to approximately solve the LPs (5) and (6) via stochastic primal-dual optimization. Throughout this section, we will assume sampling access to ν_0 and a generative model of the MDP that can produce i.i.d. samples from $P(\cdot|x, a)$ for any of the core state-action pairs.

We first introduce the *Lagrangian function* associated with the two LPs:

$$\mathcal{L}(\lambda, u; \theta, V) = \langle \lambda, \mathcal{U}(r + \gamma PV - Q_\theta) \rangle + (1 - \gamma)\langle \nu_0, V \rangle + \langle u, Q_\theta - EV \rangle. \tag{7}$$

To facilitate optimization, we will restrict the decision variables to some naturally chosen compact sets. For the primal variables, we require that $\lambda \in \Delta_{\widetilde{\mathcal{Z}}}$ and $u \in \Delta_{\mathcal{Z}}$. For θ, we restrict our attention to the domain $\mathbb{B}(D_\gamma) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq D_\gamma\}$ and we will consider only $V \in S = \{V \in \mathbb{R}^X \mid \|V\|_\infty \leq RD_\gamma\}$. Hence, we seek to solve

$$\min_{\theta \in \mathbb{B}(D_\gamma), V \in S} \max_{\lambda \in \Delta_{\widetilde{\mathcal{Z}}}, u \in \Delta_{\mathcal{Z}}} \mathcal{L}(\lambda, u; \theta, V).$$

Our algorithm is inspired by the classic stochastic optimization recipe of running two regret-minimization algorithms for optimizing the primal and dual variables, and in particular using two instances of mirror descent (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003) to update the low-dimensional decision variables θ and λ. There are, however, some significant changes to the basic recipe that are made necessary by the specifics of the problem that we consider. The first major challenge that we need to overcome is that the variables u and V are high-dimensional, so running mirror descent on them would result in an excessively costly algorithm with runtime scaling linearly with the size of the state space. To avoid this computational burden, we design a special-purpose implicit update rule for these variables that allow an efficient implementation without having to loop over the entire state space. The second challenge is somewhat more subtle: the implicit updates employed to calculate u make it impossible to adapt the standard method of extracting a policy from the solution (Wang, 2017; Cheng et al., 2020; Jin and Sidford, 2020), which necessitates an alternative policy extraction method. This in turn requires more careful "policy evaluation" steps in the implementation, which is technically achieved by performing several dual updates on θ between each primal update to λ and u. Finally, we need to make sure that the subsequent policies calculated by the algorithm do not change too rapidly, which is addressed by using a softmax policy update.

More formally, our algorithm performs the following steps in each iteration $t = 1, 2, \ldots, T$:

1. set $\nu_t = \gamma P^\mathsf{T} \mathcal{U}^\mathsf{T} \lambda_t + (1 - \gamma)\nu_0$,
2. set $u_t = \nu_t \circ \pi_t$,

3. perform K stochastic gradient descent updates starting from θ_{t-1} on $\mathcal{L}(\lambda_t, u_t; \cdot, V_{t-1})$ and set θ_t as the average of the iterates,

4. update the action-value function as $Q_t = \Phi\theta_t$,

5. update the state-value function as $V_t(x) = \sum_a \pi_t(a|x)Q_t(x,a)$.

6. perform a stochastic mirror ascent update starting from λ_t on $\mathcal{L}(\cdot, u_t; \theta_t, V_t)$ to obtain λ_{t+1},

7. update π_{t+1} as $\pi_{t+1}(a|x) \propto \pi_t(a|x)e^{\beta Q_t(x,a)}$,

We highlight that several of the above abstractly defined steps are only performed implicitly by the algorithm. First, the state distribution ν_t is never actually calculated by the algorithm, as it is only needed to generate samples for the computation of stochastic gradients with respect to θ. Note that ν_t is chosen to satisfy the primal constraint on u_t exactly. Second, the value functions V_t do not have to be calculated for all states, only for the ones being accessed by the stochastic gradient updates for λ. Similarly, the policy π_t does not need to be computed for all states, but only locally wherever necessary. The details of the stochastic updates are clarified in the pseudocode provided as Algorithm 1; we only note here that both the primal and dual updates can be implemented efficiently via a single access of the generative model per step, making for a total of $K+1$ samples per each iteration in the outer loop. Thus, the total number of times that the algorithm queries the generative model is $T(K+1)$. The algorithm finally returns a policy π_J with the index J selected uniformly at random. This policy can be written as $\pi_J(a|x) \propto \pi_1(a|x)e^{\beta \sum_{t=1}^{J-1} Q_t(x,a)}$, where the exponent can be compactly represented by the d-dimensional parameter vector $\Theta_J = \sum_{t=1}^{J-1} \theta_t$.

Algorithm 1 Global planning via primal-dual stochastic optimization.

Input: Core set $\widetilde{\mathcal{Z}}$, learning rates η, β, α, initial iterates $\theta_0 \in \mathbb{B}(D_\gamma)$, $\lambda_1 \in \Delta_{\widetilde{\mathcal{Z}}}$, $\pi_1 \in \Pi$.
for $t = 1$ **to** T **do**
  Stochastic gradient descent:
  Initialize: $\theta_t^{(1)} = \theta_{t-1}$;
  for $i = 1$ **to** K **do**
    Sample $x_{0,t}^{(i)} \sim \nu_0$ and $a_{0,t}^{(i)} \sim \pi_t(\cdot|x_{0,t}^{(i)})$,
        $(x_t^{(i)}, a_t^{(i)}) \sim \lambda_t$,
        $\overline{x}_t^{(i)} \sim P(\cdot|x_t^{(i)}, a_t^{(i)})$ and $\overline{a}_t^{(i)} \sim \pi_t(\cdot|\overline{x}_t^{(i)})$;
    Compute $\widetilde{g}_\theta(t,i) = (1-\gamma)\varphi(x_{0,t}^{(i)}, a_{0,t}^{(i)}) + \gamma\varphi(\overline{x}_t^{(i)}, \overline{a}_t^{(i)}) - \varphi(x_t^{(i)}, a_t^{(i)})$;
    Update $\theta_t^{(i+1)} = \Pi_{\mathbb{B}(D_\gamma)}(\theta_t^{(i)} - \alpha\widetilde{g}_\theta(t,i))$;
  end for
  Compute $\theta_t = \frac{1}{K}\sum_{i=1}^K \theta_t^{(i)}$;
  Stochastic mirror ascent:
  Sample $(x_t, a_t) \sim \text{Unif}(\widetilde{\mathcal{Z}})$, $(r_t, y_t) \sim P(\cdot|x_t, a_t)$;
  Compute $V_t(y_t) = \sum_a \pi_t(a|y_t)Q_t(y_t, a)$;
  Compute $\widetilde{g}_\lambda(t) = m[r(x_t, a_t) + \gamma V_t(y_t) - Q_t(x_t, a_t)]e_{(x_t, a_t)}$;
  Update $\lambda_{t+1} = \lambda_t e^{\eta\widetilde{g}_\lambda(t)}/\langle\lambda_t e^{\eta\widetilde{g}_\lambda(t)}, \mathbf{1}\rangle$;
  Policy update:
  Compute $\pi_{t+1} = \text{softmax}\left(\beta\sum_{k=1}^t \Phi\theta_k\right)$;
end for
Return: π_J with $J \sim \text{Unif}\{1, \cdots, T\}$.

Our main result regarding the performance the algorithm is the following.

Theorem 4 *Suppose that Assumption 1 holds and that the initial policy π_1 is uniform over the actions. Then, after T iterations Algorithm 1 outputs a policy π_{out} satisfying*

$$\mathbb{E}\left[\langle \mu^* - \mu^{\pi_{out}}, r\rangle\right] \leq \varepsilon_{\mathrm{opt}} + \varepsilon_{\mathrm{approx}}.$$

Here, $\varepsilon_{\mathrm{opt}}$ is a bound on the expected optimization error defined as

$$\varepsilon_{\mathrm{opt}} = \frac{\mathcal{D}\left(\lambda^* \| \lambda_1\right)}{\eta T} + \frac{\log|\mathcal{A}|}{\beta T} + \frac{2D_\gamma^2}{\alpha K} + \frac{\eta m^2(1 + 2RD_\gamma)^2}{2} + \frac{\beta R^2 D_\gamma^2}{2} + 2\alpha R^2,$$

and $\varepsilon_{\mathrm{approx}}$ is a bound on the expected approximation error defined as

$$\varepsilon_{\mathrm{approx}} = 2\mathbb{E}\left[\varepsilon_{\pi_{out}}\right] + 2\mathrm{IBE} + 2D_\gamma\left\langle \mu^*, \varepsilon_{\mathrm{core}}\right\rangle.$$

In particular, for any target accuracy $\varepsilon > 0$, setting $K = T/\left(m^2 \log\left(m|\mathcal{A}|\right)\right)$ and tuning the hyperparameters appropriately, the expected optimization error satisfies $\varepsilon_{\mathrm{opt}} \leq \varepsilon$ after n_ε queries to the generative model with

$$n_\varepsilon = \mathcal{O}\left(\frac{m^2 R^4 D_\gamma^4 \log(m|\mathcal{A}|)}{\varepsilon^4}\right).$$

A few comments are in order. First, recall that the Q-approximation error $\varepsilon_{\pi_{out}}$ and the inherent Bellman error terms are zero for linear MDPs, where the only remaining approximation error term corresponds to the extent of violation of the core state-action assumption. Interestingly, our result shows that this approximation error does not need to be uniformly small, but only needs to be under control in the states that the optimal policy visits. Thus, provided access to good core state-action pairs, our algorithm is guaranteed to output a near-optimal policy after polynomially many queries to the generative model. Furthermore, we note that the computational complexity of our algorithm exactly matches its sample complexity up to a factor of the number of actions, which is the cost of sampling from the softmax policies. To see why this is the case, note that for each sample drawn from the simulator, the initial-state distribution and the softmax policy, the algorithm performs a constant number of elementary operations. Finally, we once again stress that our algorithm outputs a globally valid softmax policy that is compactly represented by a d-dimensional parameter vector. Thus, to our knowledge, our method is the first global planning method that produces a simple output while being provably efficient both statistically and computationally under a linear MDP assumption (and even relaxed versions thereof).

5. Analysis

This section presents the main components of the proof of our main result, Theorem 4. The analysis relies on the definition of a quantity we call the *dynamic duality gap* associated with the iterates of the algorithm, defined with respect to any *primal comparator* (λ^*, u^*) and *dual comparator sequence* $\theta_{1:T}^* = (\theta_1^*, \ldots, \theta_T^*)$ and $V_{1:T}^* = (V_1^*, \ldots, V_T^*)$ as

$$\mathcal{G}_T\left(\lambda^*, u^*; \theta_{1:T}^*, V_{1:T}^*\right) = \frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{L}(\lambda^*, u^*; \theta_t, V_t) - \mathcal{L}(\lambda_t, u_t; \theta_t^*, V_t^*)\right).$$

The dynamic duality gap is closely related to the classic notion of duality gap considered in the saddle-point-optimization literature, with the key difference being that the dual comparator is not a static point (θ^*, V^*), but is rather a sequence of comparator points. Similarly to how the duality gap can be written as the sum of the average regrets of two concurrent regret minimization methods for the primal and dual method, the dynamic duality gap can be written as the sum of the average regret of the primal method and the *dynamic* regret of the dual method. In our analysis below, we relate the dynamic duality gap to the expected suboptimality of the policy π_{out} produced by our algorithm, and show how the dynamic duality gap itself can be bounded.

Our first lemma shows that the dynamic duality gap evaluated at an appropriately selected comparator sequence can be exactly related to the quantity of our main interest:

Lemma 5 *Let μ^* denote the occupancy measure of an optimal policy and $\lambda^* = \mathcal{B}^\mathsf{T}\mu^*$. Also, let $V_t^* = V^{\pi_t}$ and $\theta_t^* = \arg\min_{\theta \in \mathbb{B}(D_\gamma)} \|Q^{\pi_t} - \Phi\theta\|_\infty$. Then,*

$$\mathcal{G}_T\left(\lambda^*, \mu^*; \theta_{1:T}^*, V_{1:T}^*\right) \geq \mathbb{E}_T\left[\langle \mu^* - \mu^{\pi_{out}}, r\rangle\right] - 2\mathbb{E}_T\left[\varepsilon_{\pi_{out}}\right] - 2\text{IBE} - D_\gamma \langle \mu^*, \varepsilon_{\text{core}}\rangle.$$

Notably, when $\varepsilon_{\text{core}} = 0$, IBE $= 0$ and $\varepsilon_{\pi_t} = 0$ hold for all t, the claim holds with equality. The proof of this result draws inspiration from Cheng et al. (2020), who first introduced the idea of making use of an adaptively chosen comparator point to reduce duality-gap guarantees to policy suboptimality guarantees in their Proposition 4 (at least to our knowledge). The idea of using a dynamic comparator sequence is new to our analysis and allows us to output a simple softmax policy that is compactly represented by a linearly parametrized Q-function. Below, we prove the lemma for the special case where all approximation errors are zero, and we relegate the slightly more complicated proof for the general case to Appendix B.4.

Proof for zero approximation error. Suppose that IBE $= 0$, $\varepsilon_{\pi_t} = 0$ holds for all t and $\Delta_{\text{core}} = 0$. Let $Q_t^* = \Phi\theta_t^* = Q^{\pi_t}$ and θ_t' be such that $\Phi\theta_t' = r + \gamma PV_t - Q_t$. We start by rewriting each term in the definition of the dynamic duality gap. First, we note that

$$\begin{aligned}
\mathcal{L}(\lambda^*, u^*; \theta_t, V_t) &= \langle \mathcal{B}^\mathsf{T}\mu^*, \mathcal{U}(r + \gamma PV_t - Q_t)\rangle + (1-\gamma)\langle \nu_0, V_t\rangle + \langle \mu^*, Q_t - EV_t\rangle \\
&= \langle \mathcal{U}^\mathsf{T}\mathcal{B}^\mathsf{T}\mu^*, r + \gamma PV_t - Q_t\rangle + \langle \mu^*, Q_t - \gamma PV_t\rangle \\
&= \langle \mathcal{U}^\mathsf{T}\mathcal{B}^\mathsf{T}\mu^*, \Phi\theta_t'\rangle - \langle \mu^*, \Phi\theta_t' - r\rangle = \langle \mu^*, r\rangle,
\end{aligned}$$

where in the last line we have used Assumption 1 with $\Delta_{\text{core}} = 0$ that implies $\mathcal{B}\mathcal{U}\Phi = \Phi$.

On the other hand, we have

$$\begin{aligned}
\mathcal{L}(\lambda_t, u_t; \theta_t^*, V_t^*) &= \langle \lambda_t, \mathcal{U}(r + \gamma PV^{\pi_t} - Q^{\pi_t})\rangle + (1-\gamma)\langle \nu_0, V^{\pi_t}\rangle + \langle u_t, Q^{\pi_t} - EV^{\pi_t}\rangle \\
&= (1-\gamma)\langle \nu_0, V^{\pi_t}\rangle = \langle \mu^{\pi_t}, r\rangle,
\end{aligned}$$

where the last step follows from the definitions of the value function and the discounted occupancy measure, and the previous step from the fact that the value functions satisfy the Bellman equations $Q^{\pi_t} = r + \gamma PV^{\pi_t}$, and that $\langle u_t, EV^{\pi_t}\rangle = \langle u_t, Q^{\pi_t}\rangle$. Indeed, this last part follows from writing

$$\langle u_t, Q^{\pi_t}\rangle = \sum_x \nu_t(x)\sum_a \pi_t(a|x)Q^{\pi_t}(x,a) = \sum_x \nu_t(x)V^{\pi_t}(x) = \langle u_t, EV^{\pi_t}\rangle,$$

where we made use of the relation between the action-value function and the value function of π_t.

Putting the above calculations together, we can conclude that

$$\mathcal{G}_T\left(\lambda^*, \mu^*; \theta_{1:T}^*, V_{1:T}^*\right) = \frac{1}{T}\sum_{t=1}^{T}\langle \mu^* - \mu^{\pi_t}, r\rangle = \mathbb{E}_T\left[\langle \mu^* - \mu^{\pi_\text{out}}, r\rangle\right],$$

thus verifying the claim of the lemma. ∎

It remains to show that Algorithm 1 indeed guarantees that the dynamic duality gap is bounded. This is done in the following lemma that states a bound in terms of the *conditional relative entropy*[2] between π^* and the initial policy π_1 defined as $\mathcal{H}\left(\pi^*\|\pi_1\right) = \sum_x \nu^{\pi^*}(x)\mathcal{D}\left(\pi^*(\cdot|x)\|\pi_1(\cdot|x)\right)$.

Lemma 6 *The dynamic duality gap associated with the iterates produced by Algorithm 1 satisfies*

$$\mathbb{E}\left[\mathcal{G}_T\left(\lambda^*, u^*; \theta_{1:T}^*, V_{1:T}^*\right)\right] \leq \frac{\mathcal{D}\left(\lambda^*\|\lambda_1\right)}{\eta T} + \frac{\mathcal{H}\left(\pi^*\|\pi_1\right)}{\beta T} + \frac{2D_\gamma^2}{\alpha K} + \frac{\eta m^2(1 + 2RD_\gamma)^2}{2} + \frac{\beta R^2 D_\gamma^2}{2} + 2\alpha R^2.$$

The proof is based on decomposing the dynamic duality gap into the sum of the average regret of the primal method and the average dynamic regret of the dual method. These regrets are then bounded via a standard analysis of stochastic mirror descent methods, as well as via a specialized analysis that takes advantage of the implicitly defined updates for the variables that would be otherwise intractable for mirror descent.

Proof We start by rewriting the dynamic duality gap as follows:

$$\mathcal{G}_T\left(\lambda^*, u^*; \theta_{1:T}^*, V_{1:T}^*\right) = \frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{L}(\lambda^*, u^*; \theta_t, V_t) - \mathcal{L}(\lambda_t, u_t; \theta_t^*, V_t^*)\right)$$

$$= \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{L}(\lambda^*, u^*; \theta_t, V_t) - \mathcal{L}(\lambda_t, u_t; \theta_t, V_t)\right)}_{\mathfrak{R}_T^\text{p}(\lambda^*, u^*)} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{L}(\lambda_t, u_t; \theta_t, V_t) - \mathcal{L}(\lambda_t, u_t; \theta_t^*, V_t^*)\right)}_{\mathfrak{R}_T^\text{d}(\theta_{1:T}^*, V_{1:T}^*)},$$

where $\mathfrak{R}_T^\text{p}$ and $\mathfrak{R}_T^\text{d}$ are the regret of the primal method and the dynamic regret of the dual method, respectively. We bound the two terms separately below. Before doing so, it will be useful to introduce the following shorthand notations for the gradient of the Lagrangian with respect to λ:

$$g_\lambda(t) = \nabla_\lambda \mathcal{L}(\lambda_t, u_t; \theta_t, V_t) = \mathcal{U}\left(r + \gamma P V_t - Q_t\right).$$

First, we rewrite the primal regret as

$$\mathfrak{R}_T^\text{p}(\lambda^*, u^*) = \sum_{t=1}^{T}\left(\langle\lambda^* - \lambda_t, \mathcal{U}\left(r + \gamma P V_t - Q_t\right)\rangle + \langle u^* - u_t, Q_t - E V_t\rangle\right)$$

$$= \sum_{t=1}^{T}\left(\langle\lambda^* - \lambda_t, g_\lambda(t)\rangle + \langle u^*, Q_t - E V_t\rangle\right),$$

where second step follows from recognizing the expression of $g_\lambda(t)$ and exploiting the properties of our choice of u_t and V_t. Indeed, notice that

$$\langle u_t, Q_t \rangle = \sum_x \nu_t(x) \sum_a \pi_t(a|x) Q_t(x,a) = \sum_x \nu_t(x) V_t(x) = \langle u_t, EV_t \rangle.$$

Controlling the remaining two terms can be achieved by the standard analysis of online stochastic mirror descent (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003). We only provide the high-level arguments here and defer the technical details to Appendix B.2. Since the primal updates on λ directly use stochastic mirror ascent updates, the standard analysis applies and gives a regret bound of

$$\mathbb{E}\left[\sum_{t=1}^T \langle \lambda^* - \lambda_t, g_\lambda(t) \rangle \right] \leq \frac{\mathcal{D}(\lambda^* \| \lambda_1)}{\eta} + \frac{\eta T m^2 (1 + 2RD_\gamma)^2}{2}.$$

To see how a mirror descent analysis applies to the second term in the above bound, note that it can be rewritten as

$$\sum_{t=1}^T \langle u^*, Q_t - EV_t \rangle = \sum_x \nu^*(x) \sum_{t=1}^T \sum_a (\pi^*(a|x) - \pi_t(a|x)) Q_t(x,a).$$

For each individual x, the sum $\sum_{t=1}^T \sum_a (\pi^*(a|x) - \pi_t(a|x)) Q_t(x,a)$ can be seen as the regret in a local online learning game with rewards $Q_t(x,\cdot)$ and decision variables $\pi_t(\cdot|x)$. Thus, noting that the update rule for $\pi_t(\cdot|x)$ precisely matches an instance of mirror descent, the standard analysis applies and gives a regret bound of

$$\sum_{t=1}^T \langle u^*, Q_t - EV_t \rangle \leq \sum_x \nu^*(x) \left(\frac{\mathcal{D}(\pi^*(\cdot|x) \| \pi_1(\cdot|x))}{\beta} + \frac{\beta T R^2 D_\gamma^2}{2} \right) = \frac{\mathcal{H}(\pi^* \| \pi_1)}{\beta} + \frac{\beta T R^2 D_\gamma^2}{2}.$$

To proceed, we rewrite the dynamic regret of the dual method as

$$\mathfrak{R}_T^d(\theta_{1:T}^*, V_{1:T}^*) = \sum_{t=1}^T \left(\langle \theta_t - \theta_t^*, \Phi^\mathsf{T} u_t - \Phi^\mathsf{T} \mathcal{U}^\mathsf{T} \lambda_t \rangle + \langle V_t - V_t^*, (1-\gamma)\nu_0 + \gamma P^\mathsf{T} \mathcal{U}^\mathsf{T} \lambda_t - E^\mathsf{T} u_t \rangle \right)$$

$$= \sum_{t=1}^T \langle \theta_t - \theta_t^*, \Phi^\mathsf{T} u_t - \Phi^\mathsf{T} \mathcal{U}^\mathsf{T} \lambda_t \rangle.$$

where the last step follows from using the definition of u_t that ensures $E^\mathsf{T} u_t = \gamma P^\mathsf{T} \mathcal{U}^\mathsf{T} \lambda_t + (1-\gamma)\nu_0$, and thus that the second term in the first line is zero. We now notice that the sum is exactly the function that the dual method optimizes within the inner loop of each update, and thus it can be controlled by analyzing the performance of the averaged SGD iterate θ_t. This analysis gives the following bound, stated and proved as Lemma 10 in Appendix B.3:

$$\mathbb{E}_t \left[\langle \theta_t - \theta_t^*, \Phi^\mathsf{T} u_t - \Phi^\mathsf{T} \mathcal{U}^\mathsf{T} \lambda_t \rangle \right] \leq \frac{\| \theta_{t-1} - \theta_t^* \|_2^2}{2\alpha K} + 2\alpha R^2. \tag{8}$$

Putting all the bounds together concludes the proof. ∎

Proof of Theorem 4. The proof of the first claim follows from combining Lemmas 5 and 6. The second part follows from optimizing the constants in the upper bound, and taking into account that the total number of queries to the generative model is $T(K + 1)$. ∎

6. Discussion

In this section, we conclude by discussing the broader context of our result, the limitations of our method, and outline some directions for future work.

The role of the core set assumption. The infamous negative result of Du et al. (2020) states that only supposing approximate Q^π-realizability up to an error of ε, there exist environments for which any algorithm that finds an ε-optimal policy would require an exponential number of queries to the generative model. On the positive side, Lattimore et al. (2020) showed that this exponential blow-up in the feature dimension or effective horizon is avoidable if one only needs the policy to be $\Omega(\varepsilon\sqrt{d})$-optimal. Like Shariff and Szepesvári (2020), our main result slightly improves on this result by assuming some additional structure on the function approximator: access to a small core set of states or state-action pairs that can realize all the other possible feature realizations as convex combinations. As shown in our Theorem 4 (and Theorem 3 of Shariff and Szepesvári (2020), this condition indeed allows sidestepping the negative results mentioned above, allowing us to achieve $O(\varepsilon)$-suboptimality with polynomial query complexity. The most burning question of all is of course how to find good core state-action pairs satisfying Assumption 1, or if there exist a small set of core state-action pairs under reasonable assumptions on the MDP and the feature map. Similar questions arise regarding previous works of Zanette et al. (2019); Shariff and Szepesvári (2020); Wang et al. (2021), but we are not aware of a reassuring answer at the moment. We note that our definition allows for approximate notions of core state-action pairs, which leaves open the possibility of incrementally growing the core set until sufficiently low error is achieved. We are optimistic that future work can realize the potential of learning core sets on the fly, based on direct interaction with the MDP rather than assuming access to a simulator, and that our results will become more broadly applicable.

The tightness of our bounds. The bounds stated in Theorem 4 are unlikely to be tight in general, as a comparison with the simple tabular case reveals. Indeed, the tabular case can be modeled in our setting with $m = d = |\mathcal{X}||\mathcal{A}|$, and thus the classic lower bounds of Azar et al. (2013) imply a query-complexity lower bound of $\Omega\left(\frac{d}{(1-\gamma)^2 \varepsilon^2}\right)$ for finding an ε-optimal policy. In contrast, our bounds are of the considerably higher order $O\left(\frac{d^2 \log(d|A|)}{(1-\gamma)^4 \varepsilon^4}\right)$. This excess complexity is largely due to the double-loop structure of our algorithm, which ensures that the dynamic regret of the dual player remains under control. We conjecture that it may be possible to avoid the double-loop structure of our method via more sophisticated algorithms incorporating techniques like two-timescale updates (Borkar, 1997) or optimistic mirror descent steps (Korpelevich, 1976; Rakhlin and Sridharan, 2013). We leave the investigation of such ideas for future work.

Realizability assumptions. Our current analysis requires realizability assumptions that are only mildly weaker than the (arguably rather strong) linear MDP assumption. Whether or not these assumptions can be significantly relaxed is an open question, but we believe that they may be necessary if one aims to output a single compactly parametrized policy as we do. Our algorithm and analysis certainly require this assumption, at least due to the intuition that our method is more

"policy-iteration-like" than "value-iteration-like" due to the extensive policy evaluation steps that it features. Still, the connection with approximate policy iteration methods is weaker than in the case of other algorithms like POLITEX (Lazic et al., 2019) or PC-PG (Agarwal et al., 2020a), which leaves some hope for relaxing our realizability assumptions.

Potential extensions. The method we develop here is potentially applicable for a much broader range of settings than planning with a generative model. In particular, we are confident that our techniques can be applicable for optimistic exploration in infinite-horizon MDPs via a combination of tools developed by Neu and Pike-Burke (2020) and Wei et al. (2021), to off-policy learning (Uehara et al., 2020; Zhan et al., 2022), or to imitation learning (Kamoutsi et al., 2021; Viano et al., 2022). We

Acknowledgments

References

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33:13399–13412, 2020a.

Alekh Agarwal, Sham Kakade, and Lin F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020b.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

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

Joan Bas-Serrano and Gergely Neu. Faster saddle-point optimization for solving large-scale Markov decision processes. In *Learning for Dynamics and Control*, pages 413–423, 2020.

Joan Bas-Serrano, Sebastian Curi, Andreas Krause, and Gergely Neu. Logistic Q-Learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3610–3618, 2021.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

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

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

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

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

Yichen Chen, Lihong Li, and Mengdi Wang. Scalable bilinear pi learning using state and action features. In *International Conference on Machine Learning*, pages 834–843. PMLR, 2018.

Ching-An Cheng, Remi Tachet Combes, Byron Boots, and Geoff Gordon. A reduction from reinforcement learning to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3514–3524, 2020.

Daniela Pucci De Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations research*, 51(6):850–865, 2003.

Daniela Pucci De Farias and Benjamin Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of operations research*, 29(3): 462–478, 2004.

Eric V. Denardo. On linear programming in a Markov decision problem. *Management Science*, 16 (5):281–288, 1970.

Francois d'Epenoux. A probabilistic production and inventory problem. *Management Science*, 10 (1):98–108, 1963.

Simon S. Du, Sham M Kakade, Ruosong Wang, and Lin F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

Yujia Jin and Aaron Sidford. Efficiently solving MDPs with stochastic mirror descent. In *International Conference on Machine Learning*, pages 4890–4900, 2020.

Angeliki Kamoutsi, Goran Banjac, and John Lygeros. Efficient performance bounds for primal-dual reinforcement learning from demonstrations. In *International Conference on Machine Learning*, pages 5257–5268, 2021.

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

Chandrashekar Lakshminarayanan, Shalabh Bhatnagar, and Csaba Szepesvári. A linearly relaxed approximate linear program for Markov decision processes. *IEEE Transactions on Automatic control*, 63(4):1185–1191, 2017.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

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

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

Prashant Mehta and Sean Meyn. Q-learning and Pontryagin's minimum principle. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 3598–3605. IEEE, 2009.

Arkadi Nemirovski and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1392–1403, 2020.

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

Marek Petrik and Shlomo Zilberstein. Constraint relaxation in approximate linear programs. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 809–816, 2009.

Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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

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

Roshan Shariff and Csaba Szepesvári. Efficient planning in large MDPs with weak linear function approximation. *arXiv preprint arXiv:2007.06184*, 2020.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31, 2018.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Daniil Tiapkin and Alexander Gasnikov. Primal-dual stochastic mirror descent for MDPs. In *International Conference on Artificial Intelligence and Statistics*, pages 9723–9740. PMLR, 2022.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 9659–9668, 2020.

Luca Viano, Angeliki Kamoutsi, Gergely Neu, Igor Krawczuk, and Volkan Cevher. Proximal point imitation learning. *arXiv preprint arXiv:2209.10968*, 2022.

Bingyan Wang, Yuling Yan, and Jianqing Fan. Sample-efficient reinforcement learning for linearly-parameterized MDPs with a generative model. *Advances in Neural Information Processing Systems*, 34, 2021.

Mengdi Wang. Primal-dual pi-learning: Sample complexity and sublinear run time for ergodic Markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

Mengdi Wang and Yichen Chen. An online primal-dual method for discounted Markov decision processes. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 4516–4521, 2016.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward MDPs with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.

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

Dong Yin, Botao Hao, Yasin Abbasi-Yadkori, Nevena Lazić, and Csaba Szepesvári. Efficient local planning with linear function approximation. In *International Conference on Algorithmic Learning Theory*, pages 1165–1192, 2022.

Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Limiting extrapolation in linear approximate value iteration. *Advances in Neural Information Processing Systems*, 32, 2019.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman error. In *International Conference on Machine Learning*, pages 10978–10989, 2020.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

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

Appendix A. Missing proofs for Section 3

A.1. Analysis of the Relaxed LP for Linear MDPs

Here we show that the optimal solutions of the original linear program are also optimal in the relaxed LP when there exist $W \in \mathbb{R}^{d \times X}$, $\vartheta \in \mathbb{R}^d$ such that $P = \Phi W$, $r = \Phi \vartheta$ and $\Delta_{\text{core}} = \mathbf{0}$. For ease of reference, we denote by $\mathcal{M}$ and $\widetilde{\mathcal{M}}_\Phi$ the feasible sets of the original LP and the relaxed LP respectively, and we let $\widetilde{\mathcal{M}}_\Phi\big|_u$ denote the set of distributions u such that $(\lambda, u) \in \widetilde{\mathcal{M}}_\Phi$ holds for some λ. Recall that $\mathcal{M}$ is the set of valid occupancy measures.

For the primal relaxed LP, using $P = \Phi W$ and $\Phi = \mathcal{B}\mathcal{U}\Phi$ it is easy to check that for all $(\lambda, u) \in \widetilde{\mathcal{M}}_\Phi$, u is a valid occupancy measure satisfying $E^\mathsf{T} u = (1 - \gamma)\nu_0 + \gamma P^\mathsf{T} u$. Furthermore, for $u \in \mathcal{M}$, choose $\lambda = \mathcal{B}^\mathsf{T} u$ so that $\Phi^\mathsf{T}\mathcal{U}^\mathsf{T}\lambda = \Phi^\mathsf{T} u$ holds, which implies feasibility as $(\lambda, u) \in \widetilde{\mathcal{M}}_\Phi$. Therefore, we have $\mathcal{M} \subseteq \widetilde{\mathcal{M}}_\Phi\big|_u$. Let $(\lambda^*, u^*) = \arg\max\{\langle \lambda, \mathcal{U}r \rangle | E^\mathsf{T} u = (1 - \gamma)\nu_0 + \gamma P^\mathsf{T}\mathcal{U}^\mathsf{T}\lambda, \Phi^\mathsf{T}\mathcal{U}^\mathsf{T}\lambda = \Phi^\mathsf{T} u, \lambda \in \mathbb{R}_+^m, u \in \mathbb{R}_+^{XA}\}$. We already know by the relation $\mathcal{M} \subseteq \widetilde{\mathcal{M}}_\Phi\big|_u$ that $\langle \mathcal{B}^\mathsf{T}\mu^*, \mathcal{U}r \rangle \leq \langle \lambda^*, \mathcal{U}r \rangle$ but since $r = \Phi \vartheta$ and $\Phi = \mathcal{B}\mathcal{U}\Phi$, we also have

$$\langle \lambda^*, \mathcal{U}r \rangle = \langle \lambda^*, \mathcal{U}\Phi\varphi \rangle = \langle \Phi^\mathsf{T}\mathcal{U}^\mathsf{T}\lambda^*, \vartheta \rangle = \langle \Phi^\mathsf{T} u^*, \vartheta \rangle = \langle u^*, \Phi\vartheta \rangle = \langle u^*, r \rangle.$$

Also

$$\langle \mathcal{B}^\mathsf{T}\mu^*, \mathcal{U}r \rangle = \langle \mu^*, \mathcal{B}\mathcal{U}r \rangle = \langle \mu^*, \mathcal{B}\mathcal{U}\Phi\vartheta \rangle = \langle \mu^*, \Phi\vartheta \rangle = \langle \mu^*, r \rangle$$

That is, $\langle \mu^*, r \rangle \leq \langle u^*, r \rangle$. But since u^* is a valid occupancy measure, we must have $\langle \mu^*, r \rangle = \langle u^*, r \rangle$ due to optimality of μ^* on the space of occupancy measures. Therefore, any optimal occupancy measure is feasible and optimal in the primal relaxed ALP when the MDP is linear and the core state action assumption is satisfied exactly.

For the dual, notice that linearity of the transition probability and reward functions imply that there exists $\theta^* \in \mathbb{R}^d$ such that $Q^* = \Phi\theta^*$. Hence, feasibility of (V^*, θ^*) follows by definition so we only need to check optimality in the relaxed dual LP. Let $T_V^* : \mathbb{R}^X \to \mathbb{R}^X$ denote the Bellman optimality operator for state value functions. We know that for any feasible (V, θ), we have $\mathcal{U}Q_\theta \geq \mathcal{U}[r + \gamma PV]$, which by monotonicity of $\mathcal{B}$ implies $\mathcal{B}\mathcal{U}Q_\theta \geq \mathcal{B}\mathcal{U}[r + \gamma PV]$. Recalling that both Q_θ and $r + \gamma PV$ are linear[3] in Φ and $\mathcal{B}\mathcal{U}\Phi = \Phi$, this further implies $Q_\theta \geq r + \gamma PV$. Then by the first constraint, we have $EV \geq Q_\theta \geq r + \gamma PV$, which implies $V \geq M[r + \gamma PV] = T_V^*V$. Then, by monotonicity of the Bellman optimality operator with V^* as its unique fixed point, for any feasible (V, θ) of the relaxed dual LP, we can repeat the above argument and obtain

$$V \geq T_V^*V \geq (T_V^*)^2 V \geq \cdots \geq V^*.$$

In a similar way using the Bellman optimality operator acting on Q-functions, we can also show that $Q_\theta \geq Q_{\theta^*}$. Therefore, (V^*, θ^*) is optimal in the relaxed LP, and these solutions are unique when ν_0 has full support.

Appendix B. Missing proofs for Section 5

B.1. Properties of the gradient estimators

Here we state a pair of results that establish some useful properties of the gradient estimators calculated by Algorithm 1. The first one concerns the gradient estimators used by the primal method.

---

3. This is the only point where the proof uses the assumption that the transition model is linear. Also, when the reward vector is known one can easily prove linearity of rewards in the feature map. Notice that this implies that the result holds under the much weaker condition of Φ being complete under the Bellman optimality operator.

Lemma 7 *The gradient estimator $\widetilde{g}_\lambda(t)$ satisfies $\mathbb{E}\left[\widetilde{g}_\lambda(t)\,|\mathcal{F}_{t-1},\theta_t\right] = \nabla_\lambda \mathcal{L}(\lambda_t, u_t; \theta_t, V_t)$ and $\|\widetilde{g}_\lambda(t)\|_\infty \leq m(1 + (1+\gamma)RD_\gamma)$.*

Proof The first claim follows from the following straightforward calculation:

$$
\begin{aligned}
\mathbb{E}\left[\widetilde{g}_\lambda(t)\,|\mathcal{F}_{t-1},\theta_t\right] &= \mathbb{E}\left[m[r(x_t,a_t) + \gamma V_t(y_t) - Q_t(x_t,a_t)]e_{(x_t,a_t)}\,|\mathcal{F}_{t-1},\theta_t\right] \\
&= m \sum_{(x,a)\in\widetilde{\mathcal{Z}}} \frac{1}{m} \sum_{y\in\mathcal{X}} P(y|x,a)[r(x,a) + \gamma V_t(y) - Q_t(x,a)]e_{(x,a)} \\
&= \sum_{(x,a)\in\widetilde{\mathcal{Z}}} \left[r(x,a) + \gamma \sum_{y\in\mathcal{X}} P(y|x,a)V_t(y) - Q_t(x,a)\right]e_{(x,a)} \\
&= \mathcal{U}[r + \gamma PV_t - Q_t] = \nabla_\lambda \mathcal{L}(\lambda_t, u_t; \theta_t, V_t).
\end{aligned}
$$

For the second half, recall that $r \in [0,1]$ and $\|V_t\|_\infty \leq \|Q_t\|_\infty \leq RD_\gamma$, so that we can write

$$
\begin{aligned}
\|\widetilde{g}_\lambda(t)\|_\infty &= \left\|m\left(r(x_t,a_t) + \gamma V_t(y_t) - Q_t(x_t,a_t)\right)e_{(x_t,a_t)}\right\|_\infty \\
&= m\left\|\left(r(x_t,a_t) + \gamma V_t(y_t) - Q_t(x_t,a_t)\right)e_{(x_t,a_t)}\right\|_\infty \\
&= m\left(r(x_t,a_t) + \gamma V_t(y_t) - Q_t(x_t,a_t)\right) \leq m(1 + (1+\gamma)RD_\gamma).
\end{aligned}
$$

This completes the proof. ■

The second result concerns the gradient estimators used by the dual method for updating θ.

Lemma 8 *The gradient estimator $\widetilde{g}_\theta(t,i)$ satisfies[4] $\mathbb{E}_{t,i}\left[\widetilde{g}_\theta(t,i)\right] = \nabla_\theta \mathcal{L}(\lambda_t, u_t; \theta_{t-1}, V_{t-1})$ and $\|\widetilde{g}_\theta(t,i)\|_2 \leq 2R$.*

Proof We first write

$$
\begin{aligned}
\mathbb{E}_{t,i}\left[\widetilde{g}_\theta(t,i)\right] &= \mathbb{E}_{t,i}\left[(1-\gamma)\varphi(x_{0,t}^{(i)}, a_{0,t}^{(i)}) + \gamma\varphi(\overline{x}_t^{(i)}, \overline{a}_t^{(i)}) - \varphi(x_t^{(i)}, a_t^{(i)})\right] \\
&= \Phi^{\mathsf{T}}\mathbb{E}\left[(1-\gamma)e_{(x_{0,t}^{(i)}, a_{0,t}^{(i)})} + \gamma e_{(\overline{x}_t^{(i)}, \overline{a}_t^{(i)})} - e_{(x_t^{(i)}, a_t^{(i)})}\,|\mathcal{F}_{t,i}\right] \\
&= (1-\gamma)\Phi^{\mathsf{T}} \sum_{(x,a)\in\mathcal{Z}} \nu_0(x)\pi_t(a|x)e_{(x,a)} \\
&\quad + \gamma\Phi^{\mathsf{T}} \sum_{(x,a)\in\mathcal{Z}} \pi_t(a|x)\left(\sum_{(x',a')\in\widetilde{\mathcal{Z}}} P(x|x',a')\lambda_t(x',a')\right)e_{(x,a)} \\
&\quad - \Phi^{\mathsf{T}} \sum_{(x,a)\in\widetilde{\mathcal{Z}}} \lambda_t(x,a)e_{(x,a)} \\
&= \Phi^{\mathsf{T}}\left(((1-\gamma)\nu_0 + \gamma P^{\mathsf{T}}\mathcal{U}^{\mathsf{T}}\lambda_t) \circ \pi_t - \mathcal{U}^{\mathsf{T}}\lambda_t\right) \\
&= \Phi^{\mathsf{T}}u_t - \Phi^{\mathsf{T}}\mathcal{U}^{\mathsf{T}}\lambda_t \\
&= \nabla_\theta \mathcal{L}(\lambda_t, u_t; \theta_{t-1}, V_{t-1})
\end{aligned}
$$

---

4. The notation $\mathbb{E}_{t,i}\left[\cdot\right]$ is as defined in the proof of Lemma 6.

Also, using that $\|\varphi(x,a)\|_2 \le R$ for all $(x,a) \in \mathcal{Z}$, we get

$$
\begin{aligned}
\|\widetilde{g}_\theta(t,i)\|_2 &= \|(1-\gamma)\varphi(x_{0,t}^{(i)}, a_{0,t}^{(i)}) + \gamma\varphi(\overline{x}_t^{(i)}, \overline{a}_t^{(i)}) - \varphi(x_t^{(i)}, a_t^{(i)})\|_2 \\
&\le (1-\gamma)\|\varphi(x_{0,t}^{(i)}, a_{0,t}^{(i)})\|_2 + \gamma\|\varphi(\overline{x}_t^{(i)}, \overline{a}_t^{(i)})\|_2 + \|\varphi(x_t^{(i)}, a_t^{(i)})\|_2 \\
&\le (1-\gamma)R + \gamma R + R = 2R,
\end{aligned}
$$

thus concluding the proof. ∎

B.2. Stochastic mirror ascent analysis

This section presents the final steps in the primal regret analysis deferred from the proof of Lemma 6. In particular, we will show an upper bound on

$$
\mathfrak{R}_T^{\mathrm{p}}(\lambda^*, u^*) = \sum_{t=1}^T \big(\langle \lambda^* - \lambda_t, g_\lambda(t) \rangle + \langle u^*, Q_t - EV_t \rangle \big).
$$

The first sum above is clearly the regret of online stochastic mirror ascent, and the second term can be written as an average of local regrets in each state x as

$$
\sum_{t=1}^T \langle u^*, Q_t - EV_t \rangle = \sum_x \nu^*(x) \sum_{t=1}^T \sum_a (\pi^*(a|x) - \pi_t(a|x)) Q_t(x,a).
$$

The following lemma will allow us to treat the two sums in a unified manner.

Lemma 9 *Given $\omega_1 \in \Delta_N$, define the sequence of vectors $\omega_2, \ldots, \omega_{n+1}$ and $g_1, g_2, \ldots, g_n$ in $\mathbb{R}^N$ via the following recursion. For each k, let $\widetilde{g}_k$ satisfy $\mathbb{E}[\widetilde{g}_k | \mathcal{F}_{k-1}] = g_k$ and let*

$$
\omega_{k+1,i} = \frac{\omega_{k,i} e^{\tau \widetilde{g}_{k,i}}}{\sum_{j=1}^N \omega_{k,j} e^{\tau \widetilde{g}_{k,j}}} \qquad \textit{for } i = 1, \cdots, N.
$$

Supposing that $\|\widetilde{g}_k\|_\infty^2 \le G^2$ holds, the following bound is satisfied for any $\omega^* \in \Delta_N$,

$$
\mathbb{E}\left[\sum_{k=1}^n \langle \omega^* - \omega_k, g_k \rangle \right] \le \frac{\mathcal{D}(\omega^* \| \omega_1)}{\tau} + \frac{\tau n G^2}{2}.
$$

We provide the standard proof below. To put this result to work, note that the recursion defining λ clearly matches the above requirements using the unbiased gradient estimators $\widetilde{g}_\lambda(t)$ that satisfy $\|\widetilde{g}_\lambda(t)\|_\infty \le m(1 + 2RD_\gamma)$ (as shown Lemma 7 in Appendix B.1). Thus, we can apply Lemma 9 to obtain the bound

$$
\mathbb{E}\left[\sum_{t=1}^T \langle \lambda^* - \lambda_t, g_\lambda(t) \rangle \right] \le \frac{\mathcal{D}(\lambda^* \| \lambda_1)}{\eta} + \frac{\eta T m^2 (1 + 2RD_\gamma)^2}{2}.
$$

In order to use the result again to bound the other term, notice that the sequence of policies $\pi_t(\cdot|x)$ and action-value functions $Q_t(x, \cdot)$ are indeed generated by a recursion of the form required by

Lemma 9. Further noting that $\|Q_t\|_\infty \leq RD_\gamma$, we can apply the lemma to each x individually to obtain

$$\sum_{t=1}^{T} \langle u^*, Q_t - EV_t \rangle \leq \sum_x \nu^*(x) \left(\frac{\mathcal{D}\left(\pi^*(\cdot|x)\|\pi_1(\cdot|x)\right)}{\beta} + \frac{\beta TR^2D_\gamma^2}{2} \right) = \frac{\mathcal{H}\left(\pi^*\|\pi_1\right)}{\beta} + \frac{\beta TR^2D_\gamma^2}{2}.$$

This proves the claims in the proof of Lemma 6.

Proof of Lemma 9   Using that $\omega^*, \{\omega_k\}_{k=1}^n \in \Delta_N$, the proof follows from straightforward calculations involving the relative entropies between ω^* and iterates ω_k and ω_{k+1}:

$$
\begin{aligned}
\mathcal{D}\left(\omega^*\|\omega_{k+1}\right) &= \sum_{i=1}^{N} \omega_i^* \log \frac{\omega_i^*}{\omega_{k+1,i}} \\
&= \sum_{i=1}^{N} \omega_i^* \log \frac{\omega_i^*}{\omega_{k,i}} - \sum_{i=1}^{N} \omega_i^* \log \frac{\omega_{k+1,i}}{\omega_{k,i}} \\
&= \mathcal{D}\left(\omega^*\|\omega_k\right) - \sum_{i=1}^{N} \omega_i^* \log \frac{\omega_{k,i}e^{\tau\widetilde{g}_{k,i}}}{\omega_{k,i}\sum_{j=1}^{N}\omega_{k,j}e^{\tau\widetilde{g}_{k,j}}} \\
&= \mathcal{D}\left(\omega^*\|\omega_k\right) - \sum_{i=1}^{N} \omega_i^* \log \frac{e^{\tau\widetilde{g}_{k,i}}}{\sum_{j=1}^{N}\omega_{k,j}e^{\tau\widetilde{g}_{k,j}}} \\
&= \mathcal{D}\left(\omega^*\|\omega_k\right) - \langle\omega^*, \tau\widetilde{g}_k\rangle + \log \sum_{i=1}^{N}\omega_{k,i}e^{\tau\widetilde{g}_{k,i}} \\
&= \mathcal{D}\left(\omega^*\|\omega_k\right) - \langle\omega^* - \omega_k, \tau\widetilde{g}_k\rangle - \langle\omega_k, \tau\widetilde{g}_k\rangle + \log \sum_{i=1}^{N}\omega_{k,i}e^{\tau\widetilde{g}_{k,i}} \\
&\leq \mathcal{D}\left(\omega^*\|\omega_k\right) - \langle\omega^* - \omega_k, \tau\widetilde{g}_k\rangle + \frac{\tau^2}{2}\|\widetilde{g}_k\|_\infty^2,
\end{aligned}
$$

where the last inequality follows from the condition on τ and Hoeffding's lemma (cf. Lemma A.1 in Cesa-Bianchi and Lugosi, 2006). Taking conditional expectations on both sides, using that $\mathbb{E}\left[\widetilde{g}_k|\mathcal{F}_{k-1}\right] = g_k$, and reordering the terms, we get

$$\langle\omega^* - \omega_k, g_k\rangle \leq \frac{\mathcal{D}\left(\omega^*\|\omega_k\right) - \mathbb{E}_k\left[\mathcal{D}\left(\omega^*\|\omega_{k+1}\right)\right]}{\tau} + \frac{\tau}{2}\mathbb{E}_k\left[\|\widetilde{g}_k\|_\infty^2\right].$$

After summing up for all k and taking marginal expectations on both sides, we obtain

$$\mathbb{E}\left[\sum_{k=1}^{n}\langle\omega^* - \omega_k, g_k\rangle\right] \leq \frac{\mathcal{D}\left(\omega^*\|\omega_1\right) - \mathbb{E}\left[\mathcal{D}\left(\omega^*\|\omega_{T+1}\right)\right]}{\tau} + \frac{\tau}{2}\sum_{k=1}^{n}\mathbb{E}\left[\|\widetilde{g}_k\|_\infty^2\right].$$

Upper bounding the negative divergence term by 0 and using that $\|\widetilde{g}_k\|_\infty^2 \leq G^2$ holds for $k = 1, \cdots, n$ concludes the proof. ∎

B.3. Stochastic gradient descent analysis

In this section we bound the optimization error incurred in each dual update to the parameter vector θ. Combined with the facts that $\|\theta_{t-1} - \theta_t^*\| \leq 2D_\gamma$ and using the gradient bounds stated in Lemma 8, the following lemma proves the bound of Equation (8) appearing in the main text.

Lemma 10 *For any $\theta_t^* \in \mathbb{B}(D_\gamma)$, the iterate θ_t satisfies the bound*

$$\langle \theta_t - \theta_t^*, \Phi^\mathsf{T} u_t - \Phi^\mathsf{T} \mathcal{U}^\mathsf{T} \lambda_t \rangle \leq \frac{\|\theta_{t-1} - \theta_t^*\|_2^2}{2\alpha K} + \frac{\alpha}{2K} \sum_{i=1}^{K} \mathbb{E}\left[\|\widetilde{g}_\theta(t,i)\|_2^2\right].$$

Proof The proof follows from classic calculations that can be familiar from references like Nemirovski and Yudin (1983) or Zinkevich (2003). We will use $\mathcal{F}_{t,i}$ to denote the sigma-algebra generated by the interaction history up to iteration $i-1$ within the inner loop for updating the dual variables in round t. We will use the shorthand notation $\mathbb{E}_{t,i}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t,i}]$ to signify expectations conditioned on the history up to this point. Using the definition of $\theta_t^{(i+1)}$ in Algorithm 1, the following bound holds for any $\theta_t^* \in \mathbb{B}(D_\gamma)$, i and t:

$$\begin{aligned}
\|\theta_t^{(i+1)} - \theta_t^*\|_2^2 &= \|\Pi_{\mathbb{B}(D_\gamma)}(\theta_t^{(i)} - \alpha\widetilde{g}_\theta(t,i)) - \theta_t^*\|_2^2 \\
&\leq \|\theta_t^{(i)} - \theta_t^* - \alpha\widetilde{g}_\theta(t,i)\|_2^2 \\
&= \|\theta_t^{(i)} - \theta_t^*\|_2^2 - 2\alpha\langle \theta_t^{(i)} - \theta_t^*, \widetilde{g}_\theta(t,i)\rangle + \alpha^2\|\widetilde{g}_\theta(t,i)\|_2^2,
\end{aligned}$$

where the inequality follows from the fact that the projection operator is a nonexpansion with respect to the Euclidean norm and the rest follows from straightforward manipulations. After reordering, taking conditional expectations on both sides, and observing that $\mathbb{E}_{t,i}[\widetilde{g}_\theta(t,i)] = \Phi^\mathsf{T} u_t - \Phi^\mathsf{T} \mathcal{U}^\mathsf{T} \lambda_t$, we obtain

$$\langle \theta_t^{(i)} - \theta_t^*, \Phi^\mathsf{T} u_t - \Phi^\mathsf{T} \mathcal{U}^\mathsf{T} \lambda_t \rangle \leq \frac{\|\theta_t^{(i)} - \theta_t^*\|_2^2 - \mathbb{E}_{t,i}\left[\|\theta_t^{(i+1)} - \theta_t^*\|_2^2\right]}{2\alpha} + \frac{\alpha}{2}\mathbb{E}_{t,i}\left[\|\widetilde{g}_\theta(t,i)\|_2^2\right]$$

Then, summing over i and taking the marginal expectation gives

$$\mathbb{E}\left[\sum_{i=1}^{K}\langle \theta_t^{(i)} - \theta_t^*, \Phi^\mathsf{T} u_t - \Phi^\mathsf{T} \mathcal{U}^\mathsf{T} \lambda_t \rangle\right] \leq \frac{\|\theta_t^{(1)} - \theta_t^*\|_2^2}{2\alpha} + \frac{\alpha}{2}\sum_{i=1}^{K}\mathbb{E}\left[\|\widetilde{g}_\theta(t,i)\|_2^2\right]$$

Dividing both sides by K and recalling the definition of θ_t concludes the proof. ∎

B.4. The full proof of Lemma 5

Proof Let $Q_t^* = \Phi\theta_t^*$ and $\delta_t^* = Q^{\pi_t} - Q_t^*$, and also $\theta_t' = \arg\min_{\theta \in \mathcal{B}(D_\gamma)} \langle(\mathcal{U}^\mathsf{T}\mathcal{B}^\mathsf{T} - I)\mu^*, r + \gamma P V_t - \Phi\theta\rangle$ and $\delta_t = r + \gamma P V_t - \Phi\theta_t'$. We start by rewriting each term in the definition of the dynamic duality gap. First, we consider the following term:

$$\begin{aligned}
\mathcal{L}(\lambda^*, u^*; \theta_t, V_t) &= \langle \mathcal{B}^\mathsf{T}\mu^*, \mathcal{U}(r + \gamma P V_t - Q_t)\rangle + (1-\gamma)\langle\nu_0, V_t\rangle + \langle\mu^*, Q_t - EV_t\rangle \\
&= \langle \mathcal{U}^\mathsf{T}\mathcal{B}^\mathsf{T}\mu^*, r + \gamma P V_t - Q_t\rangle + \langle\mu^*, Q_t - \gamma P V_t\rangle \\
&= \langle \mathcal{U}^\mathsf{T}\mathcal{B}^\mathsf{T}\mu^*, \Phi\left(\theta_t - \theta_t'\right) + \delta_t\rangle - \langle\mu^*, \Phi\left(\theta_t - \theta_t'\right) + \delta_t\rangle + \langle\mu^*, r\rangle \\
&= \langle\mu^*, r\rangle + \langle(\mathcal{U}^\mathsf{T}\mathcal{B}^\mathsf{T} - I)\mu^*, \delta_t\rangle + \langle\mu^*, \Delta_{\text{core}}(\theta_t - \theta_t')\rangle,
\end{aligned}$$

22

where in the last line we have used Assumption 1 that implies $\mathcal{B}\mathcal{U}\Phi = \Phi - \Delta_{\text{core}}$. The last term can be bounded as $\langle \mu^*, \Delta_{\text{core}}(\theta_t - \theta_t') \rangle \leq \langle \mu^*, \varepsilon_{\text{core}} \rangle \|\theta_t - \theta_t'\|_2 \leq 2D_\gamma \langle \mu^*, \varepsilon_{\text{core}} \rangle$. Finally, the term in the middle can be bounded as follows, using the definition of the Inherent Bellman Error (Definition 3):

$$
\begin{aligned}
\langle (\mathcal{U}^\mathsf{T}\mathcal{B}^\mathsf{T} - I)\,\mu^*, \delta_t \rangle &= \langle (\mathcal{U}^\mathsf{T}\mathcal{B}^\mathsf{T} - I)\,\mu^*, r + \gamma PV_t - \Phi\theta_t' \rangle \\
&= \langle (\mathcal{U}^\mathsf{T}\mathcal{B}^\mathsf{T} - I)\,\mu^*, \mathrm{BE}_{\pi_t}(\theta_t', \theta_t) \rangle \\
&= \min_{\theta \in \mathbb{B}(D_\gamma)} \langle (\mathcal{U}^\mathsf{T}\mathcal{B}^\mathsf{T} - I)\,\mu^*, \mathrm{BE}_{\pi_t}(\theta, \theta_t) \rangle \leq 2\mathrm{IBE}.
\end{aligned}
\tag{9}
$$

On the other hand, we have

$$
\begin{aligned}
\mathcal{L}(\lambda_t, u_t; \theta_t^*, V_t^*) &= \langle \lambda_t, \mathcal{U}\,(r + \gamma PV_t^* - Q_t^*) \rangle + (1 - \gamma)\langle \nu_0, V_t^* \rangle + \langle u_t, Q_t^* - EV_t^* \rangle \\
&= \langle \lambda_t, \mathcal{U}\,(r + \gamma PV^{\pi_t} - Q^{\pi_t} + \delta_t^*) \rangle + (1 - \gamma)\langle \nu_0, V^{\pi_t} \rangle + \langle u_t, Q^{\pi_t} - EV^{\pi_t} - \delta_t^* \rangle \\
&= \langle \mu^{\pi_t}, r \rangle + \langle \mathcal{U}^\mathsf{T}\lambda_t - u_t, \delta_t^* \rangle,
\end{aligned}
$$

where the last step follows from observing that $(1 - \gamma)\langle \nu_0, V^{\pi_t} \rangle = \langle \mu^{\pi_t}, r \rangle$ by the definition of the value function and the occupancy measure, and also that $\langle u_t, EV^{\pi_t} \rangle = \langle u_t, Q^{\pi_t} \rangle$ and that the value functions satisfy the Bellman optimality equations $Q^{\pi_t} = r + \gamma PV^{\pi_t}$. In order to bound the last term, notice that by the definition of the Q-evaluation error in Definition 2, we have $\|\delta_t^*\|_\infty \leq \varepsilon_{\pi_t}$, and thus we have $\langle \mathcal{U}^\mathsf{T}\lambda_t - u_t, \delta_t^* \rangle \leq 2\varepsilon_{\pi_t}$.

Putting the above bounds together, we can conclude that the dynamic duality gap at the chosen set of comparators can be bounded as

$$
\begin{aligned}
\mathcal{G}_T\,(\lambda^*, \mu^*; \theta_{1:T}^*, V_{1:T}^*) &\geq \frac{1}{T}\sum_{t=1}^{T} (\langle \mu^* - \mu^{\pi_t}, r \rangle - 2\varepsilon_{\pi_t}) - 2\mathrm{IBE} - 2D_\gamma \langle \mu^*, \varepsilon_{\text{core}} \rangle \\
&= \mathbb{E}\,[\langle \mu^* - \mu^{\pi_{\text{out}}}, r \rangle] - 2\mathbb{E}\,[\varepsilon_{\pi_{\text{out}}}] - 2\mathrm{IBE} - 2D_\gamma \langle \mu^*, \varepsilon_{\text{core}} \rangle.
\end{aligned}
$$

Reordering the terms completes the proof. ∎