A Sparse Linear Program for Global Planning in Large MDPs

Gergely Neu gergely.neu@gmail.com
Universitat Pompeu Fabra, Barcelona, Spain

Nneka Okolo nnekamaureen.okolo@upf.edu
Universitat Pompeu Fabra, Barcelona, Spain

Abstract

We study a linear programming (LP) approach to planning in large Markov Decision Processes (MDPs), addressing some well-known tractability issues of traditional LP-based approaches. Starting from an LP formulation involving state-action value functions originally due to Mehta and Meyn (2009), we propose a method for reducing both the number of constraints and the number of variables, and study the conditions under which a near-optimal action-value function can be extracted from the solution of the LP. Precisely, we show that whenever the optimal Q-function is nearly realizable by a set of known features, and the feature space is covered by a small number of core state-action pairs, the solution of our reduced LP will be a close approximation of the optimal Q-function. This result significantly extends previous work that only considered state-value functions, and gives hope that LP-based methods can be effective for finding globally optimal policies.

Keywords: Markov decision processes, Linear Programming, planning, linear function approximation

1. Introduction

The linear programming (LP) formulation of Markov Decision Processes (MDPs) has been widely studied in management science and operations research since the seminal works of Manne (1960); Denardo (1970); Hordijk and Kallenberg (1979). In recent years LP-based approaches have gained traction among RL theorists as a computational recipe for provably efficient, theoretically backed algorithms (Zimin and Neu, 2013; Nachum et al., 2019a,b; Neu and Pike-Burke, 2020; Uehara et al., 2020; Bas-Serrano et al., 2021).

In the discounted MDP (DMDP) setting, the conventional LP formulates the optimal control problem as a search within the space of state-value functions constrained by the Bellman optimality equation (corresponding to a dual view), or alternatively a search within the space of occupancy measures with statistical equilibrium constraints on the induced state distribution (corresponding to a primal view). Though desirable in the context of reinforcement learning (RL), celebrated results based on this setup are fairly limited to performance guarantees on "idealized policies" extracted from state value functions (De Farias and Van Roy, 2003; Lakshminarayanan et al., 2017), which rely on the rather strong assumption of full knowledge of transition probabilities. More recently, this issue has been addressed by Shariff and Szepesvári (2020) who proposed an LP-based "local planning" approach that takes as input a state and recommends an action after solving an appropriately reduced version of the standard LP. While the theoretical guarantees in this work are very promising, the approach is very computationally intense due to the need to replan from each input state. In fact, the approach of Shariff and Szepesvári (2020) is hindered by the fact that it is based on the traditional LP involving state-value functions, which leaves no room for extracting a globally near-optimal policy after a single round of computation. In this paper, we address this issue by studying a tractable alternative to the standard LP that involves action-value functions, whose solutions can be directly used to encode policies globally.

Our starting point is an LP involving action-value functions first proposed by Mehta and Meyn (2009) and later rediscovered by Neu and Pike-Burke (2020) and Bas-Serrano et al. (2021) who reduced the number of variables via linear function approximation. We extend their results by adapting a constraint reduction method first proposed by Shariff and Szepesvári. Our key contribution is proposing a natural variant of their "core state" assumption that we call "core state-action pair assumption", and showing bounds on the approximation error associated with the resulting LP under a standard weak assumption of the optimal Q-function being nearly realizable by the feature map. Our techniques draw inspiration from the classic work of De Farias and Van Roy (2003) and build on more recent progress by Lakshminarayanan et al. (2017) and Shariff and Szepesvári (2020). Ultimately, we arrive at approximation error bounds similar to those presented in these papers but for state-action value functions.

2. Preliminaries

We study a discounted Markov Decision Processes (DMDP, Puterman, 2014) denoted by the quintuple $(\mathcal{X}, \mathcal{A}, r, P, \gamma)$ with $\mathcal{X}$ and $\mathcal{A}$ representing finite (yet potentially 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 $P : \mathcal{X} \times \mathcal{A} \to \Delta_{\mathcal{X}}$ with $\Delta_{\mathcal{X}} = \{p \in \mathbb{R}_+^{|\mathcal{X}|} | \|p\|_1 = 1\}$. 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 $(Pv)(x,a) = \sum_{x'} P(x'|x,a)v(x')$. Its adjoint $P^\mathsf{T} \in \mathbb{R}^{X \times XA}$ is similarly defined on functions $u \in \mathbb{R}^{XA}$ via the assignment $(P^\mathsf{T} u)(x) = \sum_{x',a'} P(x|x',a')$. We also define the operator $E \in \mathbb{R}^{XA \times X}$ with adjoint $E^\mathsf{T} \in \mathbb{R}^{X \times XA}$ acting on respective vectors $v \in \mathbb{R}^X, u \in \mathbb{R}^{XA}$ through the assignment $(Ev)(x,a) = v(x)$, $(E^\mathsf{T} u)(x) = \sum_a u(x,a)$, for all $x \in \mathcal{X}, a \in \mathcal{A}$. 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 $t = 0, 1, \dots,$ 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 DMDP setting that we consider, the agent's goal is to maximize its normalised 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, and the potential randomness injected by the agent. Optimal policies are typically characterized via dynamic programming methods and the optimal value functions V^* and Q^*, whose standard definitions we omit here and refer the reader to Puterman (2014). Our approach is better described in terms of the following linear program due to Mehta and Meyn (2009):

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

Feasible solutions of this LP are called occupancy measures satisfying

$$E^\mathsf{T} u = (1-\gamma)\nu_0 + \gamma P^\mathsf{T} u. \tag{2}$$

Each occupancy measure u induces a policy $\pi_u \in \Delta_{\mathcal{A}|\mathcal{X}}$, with $\pi_u(a|x) = \frac{u(x,a)}{\sum_{a'} u(x,a')}$ for all states x where the denominator is not zero, and $\pi_u(\cdot|x)$ defined arbitrarily for other states. It can be shown that an agent selecting actions according to the policy as $a_t \sim \pi_u(\cdot|x_t)$ will cover the state-action space according to the occupancy measure u in the sense that $u(x,a) = (1-\gamma)\mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \mathbb{I}_{\{(x_t,a_t)=(x,a)\}}\right]$ (Puterman, 2014). An optimal occupancy measure u^* maximizing the objective of the above LP clearly induces an optimal policy π^* with maximal normalized discounted return.

The dual of the above LP is given as follows:

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

For this LP, it is easy to see that the optimal value functions $(Q,V) = (Q^*, V^*)$ are optimal, and V^* is uniquely part of each optimal solution provided ν_0 is fully supported on all states. The optimal action-value function Q^* is particularly useful in that it directly encodes optimal policies: any optimal policy only takes actions within $\arg\max_a Q(x,a)$ with positive probability. However, one must be wary of the fact that the optimal action-value function (Q^*, V^*) is not an unique optimal solution to the primal LP, and other (less useful) solutions such as $(Q,V) = (EV^*, V^*)$ exist.

These linear programs naturally extend the classical LPs originally proposed by Manne (1960); Denardo (1970); Hordijk and Kallenberg (1979) that only involve value functions V and a single set of primal variables μ. These LPs have inspired a number of developments in approximate dynamic programming that are reviewed in Section 5. One limitation that all approaches derived from these simpler LPs is that their optimal solution V^* does not directly encode optimal policies, and additional work is required to extract good policies from near-optimal solutions. This is aptly addressed by the above LPs which we use as starting point for our work.

3. A Reduced Approximate Linear Program

As pointed out in earlier works, the above LP has outstanding potential due to being able to output Q-functions, which can efficiently encode policies. Regardless of this potential, it still has the shortcoming of having decision variables and constraints whose cardinality scales at least with the number of states and actions. To remedy these challenges, we will assume that the learner has access to a feature map $\varphi : Z \to \mathbb{R}^d$ that can effectively represent the optimal Q-function. We also assume that the feature vectors of a small number of core state-action pairs can effectively represent the feature vectors of the entire state-action space. Concretely, we make the following assumptions on the feature map, using Φ to denote the $Z \times d$ feature matrix whose (x, a)-th column corresponds to $\varphi(x, a)$.

Assumption 1 *The all-ones vector* $\mathbf{1}$ *is contained in the column span of* Φ*. That is, there exists* $\vartheta \in \mathbb{R}^d$ *such that* $\langle \varphi(x, a) \,, \vartheta \rangle = 1$ *holds for all* $x \in \mathcal{X}, a \in \mathcal{A}$.

Assumption 2 *(Realizability Assumption) The feature map* Φ *is expressive enough to approximate* Q^* *up to some small approximation error* $\varepsilon \geq 0$*. That is,* $\varepsilon = \inf_{\theta \in \mathbb{R}^d} \|Q^* - \Phi\theta\|_\infty$ *is small.*

Assumption 3 *(Core State-Action Assumption) The feature vector of any state-action pair* $(x, a) \in \mathcal{Z}$ *can be expressed as a convex combination of features evaluated at core state-action pairs* $(x', a') \in \widetilde{\mathcal{Z}} \subset \mathcal{Z}$ *with* $|\widetilde{\mathcal{Z}}| = m$ *sufficiently small. That is, for each* $(x, a) \in \mathcal{Z}$*, there exists a set of coefficients* $b(x', a'|x, a) \geq 0$ *and* $\sum_{x',a'} b(x', a'|x, a) = 1$ *such that* $\varphi(x, a) = \sum_{x',a'} b(x', a'|x, a)\varphi(x', a')$.

It will be useful to rephrase the last 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. From Assumption 3, the interpolation coefficients 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, $\Phi = \mathcal{B}\mathcal{U}\Phi$. Also, $\mathcal{U}\mathbf{1} = \mathbf{1}$, $\mathcal{B}\mathbf{1} = \mathbf{1}$ and $\mathcal{B}\mathcal{U}\mathbf{1} = \mathbf{1}$ though $\mathcal{B}\mathcal{U} \neq \mathbf{I}$.

We propose Assumption 3 as a natural extension of the core state assumption of Shariff and Szepesvári (2020) for parametric functions in $\mathbb{R}^{XA}$. As pointed out by Shariff and Szepesvári, this assumption is rather strong. In our case, finding such core pairs can require checking the feature vector of all state-action pairs, and in the worst case, the number of core pairs can be as large as the number of state-action pairs. On the bright side, by an implicit result of Proposition 1 of (Zanette et al., 2019), we know that, when Q^* is well approximated at core pairs and ε small enough, the value of all state-action pairs can be interpolated from the value at core pairs with small error.

Given the assumptions above, it then becomes reasonable to optimize over the class of parametric functions of the form $Q_\theta = \Phi\theta$ while taking into consideration only constraints corresponding to core pairs. Furthermore, we exploit our knowledge of the desired result and restrict feasible pairs (μ, u) to the probability simplex $\Delta_{\mathcal{Z}}$. Plugging Q_θ into the dual LP (3) and making use of Assumption 3 that allows us consider only core state-action pairs, we obtain the following Reduced Approximate Linear Program (RALP):

$$\begin{aligned} &\max\nolimits_{\mu,u} \; \langle \mu \,, r \rangle \\ &\text{subject to } \; E^\intercal u = (1-\gamma)\nu_0 + \gamma P^\intercal \mu, \\ &\qquad\qquad \Phi^\intercal \mu = \Phi^\intercal u \,. \quad \mu \in \widetilde{\Delta}_{\mathcal{Z}}, u \in \Delta_{\mathcal{Z}} \end{aligned}$$

The dual of this LP is given as

$$\begin{aligned} &\min\nolimits_{\theta,V} \; (1-\gamma)\langle \nu_0 \,, V \rangle \\ &\text{subject to } \; EV \geq Q_\theta \,, \\ &\qquad\qquad \mathcal{U}Q_\theta \geq \mathcal{U}[r + \gamma PV] \,. \quad \theta \in \mathbb{R}^d, V \in \mathbb{R}^X \end{aligned}$$

Here, $\widetilde{\Delta}_{\mathcal{Z}} = \{\widetilde{\mu} \in \Delta_{\mathcal{Z}} \mid \widetilde{\mu}(x,a) = 0, \forall\, (x,a) \notin \widetilde{\mathcal{Z}}\}$ is the set of sparse distributions over the state-action space that are only supported on core pairs. Notably, constraint reduction via the core state-action assumption shrinks the primal search space and expands the dual search space. In particular, letting $\lambda \in \Delta_{\widetilde{\mathcal{Z}}}$ denote the Lagrange multiplier of the second dual constraint, we have that for every feasible (μ, u) of the primal RALP, $\mu = \mathcal{U}^{\intercal}\lambda \in \widetilde{\Delta}_{\mathcal{Z}} \subset \Delta_{\mathcal{Z}}$.

As desired, the RALP achieves a reduction in the number of decision variables and constraints from $(XA + X)$ and $2XA$ to $d{+}X$ and $(XA{+}m)$ while retaining the capacity to *globally* encode policies via the Q-function. We emphasize that even with a much smaller number of decision variables and constraints, Shariff and Szepesvári's reduction of the conventional LP was only sufficient for *local* planning. Their dual RALP formulated the optimal control problem as a search within the space of approximate state value functions from which one cannot extract actual policies without full knowledge of the transition probabilities. On the other hand, similar to our primal RALP, the core state assumption induced sparsity in the primal space rendering it impossible to extract reliable policies from the primal solution. In fact, their local planning strategy relied on appending the planning state to the core set so they can extract an action distribution over the planning state from the primal variable.

Notice that the number of decision variables and constraints in the reduced problem still scales at least with the number of states. Therefore, the RALP like the initial LP still raises tractability concerns. We show in 4.2 that this can be addressed by exploiting our knowledge of the optimal state value function. Furthermore, we show that the alternative RALP can yield a good approximation to Q^*

4. Approximation Error Bounds

Unlike in the case of the standard LP, it is not immediately clear that the above reduced LPs are feasible or provide reasonable solutions to the original optimal control problem. In this section, we set out to answer these questions and particularly to understand the approximation error of the optimal solutions of the reduced LPs under various models of model misspecification. We first provide results under a relaxed version of the popular "linear MDP" assumption in Section 4.1, and then provide more general results in Section 4.2.

4.1 Nearly-linear MDPs

We first consider the now-classic linear MDP assumption due to Jin et al. (2020):

Assumption 4 *(Linear MDP) Suppose that there exists $W \in \mathbb{R}^{d\times X}$, and $\delta \in \mathbb{R}^d$ such that for any $x, x' \in \mathcal{X}$, $a \in \mathcal{A}$, the transition matrix P and reward vector r can be written as:*

$$P = \Phi W, \qquad\qquad P(x'|x,a) = \langle \varphi(x,a)\,, w(x')\rangle,$$
$$r = \Phi\delta, \qquad\qquad r(x,a) = \langle \varphi(x,a)\,, \delta\rangle.$$

Let $\mathcal{M}$, $\widetilde{\mathcal{M}}_\Phi$ denote the feasible set of the primal LP (1) and RALP respectively. Recall that for any $(\mu, u) \in \mathcal{M}$, u is a valid occupancy measure and $(\mu, u) = (u^*, u^*)$ is optimal in LP (1). Under Assumption 4, earlier works Neu and Pike-Burke (2020) and Bas-Serrano et al. (2021) have shown that (u^*, u^*) is also optimal in a primal Approximate Linear Program (ALP) derived by only considering Assumption 2 (that is, using all state-action pairs as core state-action pairs). We restate their ALP formulation in Appendix A.1 for completeness. Reconciling $\widetilde{\mathcal{M}}_\Phi$ with the feasible set of this ALP, we prove that under Assumption 4, given an optimal solution $(\widetilde{u}^*, \widetilde{\mu}^*)$ of the primal RALP, $\widetilde{u}^*$ is a valid occupancy measure which attains maximum return.

Proposition 1 *Suppose that Assumption 4 holds and*

$$(\widetilde{u}^*, \widetilde{\mu}^*) = \operatorname*{arg\,max}_{(\widetilde{\mu},\widetilde{u})\in\widetilde{\mathcal{M}}_\Phi} \langle \widetilde{\mu}\,, r\rangle. \tag{4}$$

Then, $\widetilde{u}^*$ is a valid occupancy measure with $\langle \widetilde{u}^*, r\rangle = \langle u^*, r\rangle$.

In our proof of Proposition 1, we first establish a relation between the feasible set of the primal LP (1) and the primal ALP restated in Appendix A.1.

Let $\mathcal{M}$, $\mathcal{M}_\Phi$, $\widetilde{\mathcal{M}}_\Phi$ denote the feasible set of the primal LP (1), ALP and RALP respectively. Recall that $(\mu, u) = (u^*, u^*)$ is optimal in LP (1) and for any $(\mu, u) \in \mathcal{M}$, u is a valid occupancy measure satisfying Equation (2). Studying the feasible sets, $\mathcal{M}$ and $\mathcal{M}_\Phi$, it is easy to see that $\mathcal{M} \subseteq \mathcal{M}_\Phi$. Therefore $(u^*, u^*) \in \mathcal{M}_\Phi$. Introducing the linear MDP assumption, one can then show that for any $(\mu, u) \in \mathcal{M}_\Phi$, u is also a valid occupancy measure, and ultimately (u^*, u^*) is optimal in the primal ALP (Neu and Pike-Burke, 2020; Bas-Serrano et al., 2021). That is,

$$(u^*, u^*) = \underset{(\mu,u)\in\mathcal{M}_\Phi}{\arg\max}\ \langle \mu\,, r\rangle.$$

On this note, to prove Proposition 1, it suffices to study $\mathcal{M}_\Phi$ and $\widetilde{\mathcal{M}}_\Phi$.

Proof of Proposition 1 By definition of the constraint sets $\widetilde{\mathcal{M}}_\Phi$ and $\mathcal{M}_\Phi$, $\widetilde{\mathcal{M}}_\Phi \subset \mathcal{M}_\Phi$ as $\widetilde{\Delta}_\mathcal{Z} \subset \Delta_\mathcal{Z}$. That is, for all $(\widetilde{\mu}, \widetilde{u}) \in \widetilde{\mathcal{M}}_\Phi$, $(\widetilde{\mu}, \widetilde{u}) \in \mathcal{M}_\Phi$: all feasible points of the primal RALP are feasible in the primal ALP. Studying the primal RALP constraints, we also know that under Assumption 4, for any $(\widetilde{\mu}, \widetilde{u}) \in \widetilde{\mathcal{M}}_\Phi \subset \mathcal{M}_\Phi$, with $\widetilde{\mu}$ such that $\Phi^\mathsf{T}\widetilde{u} = \Phi^\mathsf{T}\widetilde{\mu}$, $\widetilde{u}$ is a valid occupancy measure satisfying Equation (2).

By the relation $\widetilde{\mathcal{M}}_\Phi \subset \mathcal{M}_\Phi$, we have

$$(\widetilde{u}^*, \widetilde{\mu}^*) = \underset{(\widetilde{\mu},\widetilde{u})\in\widetilde{\mathcal{M}}_\Phi}{\arg\max}\ \langle \widetilde{\mu}\,, r\rangle \leq \underset{(\mu,u)\in\mathcal{M}_\Phi}{\arg\max}\ \langle \mu\,, r\rangle = (u^*, u^*),$$

which implies

$$\langle \widetilde{u}^*, r\rangle \leq \langle u^*, r\rangle \tag{5}$$

On the other hand, we have

$$\begin{aligned}
\langle u^*, r\rangle &= \langle u^*, \Phi\delta\rangle && \text{(by Assumption 4)}\\
&= \langle \Phi^\mathsf{T} u^*, \delta\rangle &&\\
&= \langle \Phi^\mathsf{T} \mu^*, \delta\rangle && \text{(by constraint 2 imposed on } \mathcal{M}_\Phi)\\
&= \langle \Phi^\mathsf{T} \mathcal{U}^\mathsf{T} \mathcal{B}^\mathsf{T} \mu^*, \delta\rangle && \text{(by Assumption 3)}\\
&= \langle \Phi^\mathsf{T} \widehat{\mu}^*, \delta\rangle && \text{(with } \widehat{\mu}^* = \mathcal{U}^\mathsf{T}\mathcal{B}^\mathsf{T}\mu^* \in \widetilde{\Delta}_\mathcal{Z})\\
&\leq \langle \Phi^\mathsf{T} \widetilde{\mu}^*, \delta\rangle && \text{(by definition of } \widetilde{\mu}^* \text{ and since } (u^*, \widehat{\mu}^*) \in \widetilde{\mathcal{M}}_\Phi \text{ for } (u^*, \mu^*) \in \mathcal{M}_\Phi)\\
&= \langle \Phi^\mathsf{T} \widetilde{u}^*, \delta\rangle && \text{By condition 2 imposed in } \widetilde{\mathcal{M}}_\Phi\\
&= \langle \widetilde{u}^*, \Phi\delta\rangle &&\\
&= \langle \widetilde{u}^*, r\rangle. &&
\end{aligned}$$

Thus, the reverse inequality $\langle u^*, r\rangle \leq \langle \widetilde{u}^*, r\rangle$ also holds. Combining this with Equation (5), we have $\langle u^*, r\rangle = \langle \widetilde{u}^*, r\rangle$ and the proof is complete. ■

Next, relaxing the rather strong Linear MDP assumption of Assumption 4, we consider a scenario where the transition model is misspecified up to a constant error. We note that in this setting some solutions of the Sparse RALP may not be valid occupancy measures. Hence, is useful to understand the disparity between feasible points of the primal RALP solution and their induced state-action distribution. We study this in the following proposition and show that when the transition model is misspecified, the primal RALP solution is close enough to a valid occupancy measure if it places less mass on state-action pairs where the transition model is poorly approximated.

Proposition 2 *Suppose that the transition function can be written as $P = \Phi W + \Sigma$ for some $\Sigma \in \mathbb{R}^{Z\times X}$ and $W \in \mathbb{R}^{d\times X}$, such that $\widehat{P} = \Phi W$ is a valid transition function satisfying $\sum_{x'} \widehat{P}(x'|x,a) = 1$ and $\widehat{P}(x'|x,a) \geq 0$ for all x, a, x'. Then, for any feasible pair $(\widetilde{\mu}, \widetilde{u})$ of the primal RALP, with induced policy $\pi_{\widetilde{u}}$ and state-action distribution $\mu_{\pi_{\widetilde{u}}}$,*

$$\|\mu_{\pi_{\widetilde{u}}} - \widetilde{u}\|_1 \leq \frac{\gamma \min\{\|\Sigma^\mathsf{T}\widetilde{u}\|_1, \|\Sigma^\mathsf{T}\mu_{\pi_{\widetilde{u}}}\|_1\}}{(1-\gamma)} \tag{6}$$

Notably, the bound implies that $\widetilde{u}$ is close to its induced occupancy measure $\mu_{\pi_{\widetilde{u}}}$ if the transition function is well approximated in the state-action pairs that are highly weighted by either of these two state-action distributions. We note here that the requirement that $\widehat{P}$ be a valid transition function can be dropped at the expense of replacing the factor $\min\left\{\|\Sigma^\mathsf{T}\widetilde{u}\|_1, \|\Sigma^\mathsf{T}\mu_{\pi_{\widetilde{u}}}\|_1\right\}$ by $\|\Sigma^\mathsf{T}\widetilde{u}\|_1$. To lay a foundation for the proof, we introduce some relevant notation in Definition 3 and specify the RALP studied in the misspecified setting in Definition 4.

Definition 3 *For $u \in \mathbb{R}_+^Z$, we define the induced:*

1. *Policy:*
$$\pi_u(a|x) = \frac{u(x,a)}{\sum_{a'} u(x,a')}$$
2. *State distribution:*
$$\nu_{\pi_u}(x) = (1-\gamma)\nu_0(x) + \gamma \sum_{x'} P_{\pi_u}(x,x')\nu_{\pi_u}(x')$$
3. *Next state distribution:*
$$P_{\pi_u}(x,x') = \sum_{a'} P(x|x',a')\pi_u(a'|x')$$
4. *State-action distribution:*
$$\mu_{\pi_u}(x,a) = \pi_u(a|x)\nu_{\pi_u}(x)$$

In vector notation, $\pi_u = \frac{u}{E^\mathsf{T}u}$, $\nu_{\pi_u} = (1-\gamma)\nu_0 + \gamma P_{\pi_u}^\mathsf{T}\nu_{\pi_u}$ and $\mu_{\pi_u} = \pi_u\nu_{\pi_u}$.

Note that for $u \in \mathbb{R}_+^Z$ with $(\mu, u) \in \mathcal{M}$, $\mu_{\pi_u} = u$ holds since u is a valid occupancy measure.

Definition 4 *(Misspecified LP) Suppose that $P = \widehat{P} + \Sigma$, where $\widehat{P} = \Phi W$ represents a valid transition model. We define the corresponding* misspecified LP *as*

$$\begin{aligned} & max_{\mu,u} \ \langle \mu\,, r\rangle \\ & \textit{subject to } E^\mathsf{T}u = (1-\gamma)\nu_0 + \gamma\widehat{P}^\mathsf{T}\mu, \\ & \qquad \Phi^\mathsf{T}\mu = \Phi^\mathsf{T}u\,. \quad \mu \in \widetilde{\Delta}_\mathcal{Z}, u \in \Delta_\mathcal{Z}. \end{aligned} \tag{7}$$

We maintain the notation $\widetilde{\mathcal{M}}_\Phi$ for the feasible set of the RALP in this setting.

Now we are ready to prove Proposition 2.

Proof of Proposition 2 In the misspecified setting, since $(\widetilde{\mu}, \widetilde{u}) \in \widetilde{\mathcal{M}}_\Phi$,

$$\begin{aligned} E^\mathsf{T}\widetilde{u} &= (1-\gamma)\nu_0 + \gamma\widehat{P}^\mathsf{T}\widetilde{\mu} \\ &= (1-\gamma)\nu_0 + \gamma\widehat{P}^\mathsf{T}\widetilde{u} && \text{Since } \Phi^\mathsf{T}\widetilde{\mu} = \Phi^\mathsf{T}\widetilde{u} \\ &= (1-\gamma)\nu_0 + \gamma\widehat{P}^\mathsf{T}\widetilde{u} + \gamma\Sigma^\mathsf{T}\widetilde{u} - \gamma\Sigma^\mathsf{T}\widetilde{u} \\ &= (1-\gamma)\nu_0 + \gamma P^\mathsf{T}\widetilde{u} - \gamma\Sigma^\mathsf{T}\widetilde{u} && \text{Since } P = \widehat{P} + \Sigma \end{aligned}$$

Then, the error due to using the approximate transition model is:

$$E^\mathsf{T}\widetilde{u} - [(1-\gamma)\nu_0 + \gamma P^\mathsf{T}\widetilde{u}] = -\gamma\Sigma^\mathsf{T}\widetilde{u}$$

Also, we can write $\widetilde{u} = \pi_{\widetilde{u}}\nu_{\widetilde{u}}$ with $\nu_{\widetilde{u}} = (1-\gamma)\nu_0 + \gamma P_{\pi_{\widetilde{u}}}^\mathsf{T}\nu_{\widetilde{u}} - \gamma\Sigma^\mathsf{T}\widetilde{u}$ and induced policy $\pi_{\widetilde{u}}$, such that $E^\mathsf{T}\widetilde{u} = \nu_{\widetilde{u}}$. Then,

$$\begin{aligned} \|\mu_{\pi_{\widetilde{u}}} - \widetilde{u}\|_1 &= \|\pi_{\widetilde{u}}\nu_{\pi_{\widetilde{u}}} - \pi_{\widetilde{u}}\nu_{\widetilde{u}}\|_1 \\ &= \|\pi_{\widetilde{u}}(\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}})\|_1 \\ &= \|\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}}\|_1 \end{aligned}$$

But,

$$\begin{aligned}
\|\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}}\|_1 &= \|(1-\gamma)\nu_0 + \gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\pi_{\widetilde{u}}} - [(1-\gamma)\nu_0 + \gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\widetilde{u}} - \gamma\Sigma^{\mathsf{T}}\widetilde{u}]\|_1 \\
&= \|(1-\gamma)\nu_0 + \gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\pi_{\widetilde{u}}} - (1-\gamma)\nu_0 - \gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\widetilde{u}} + \gamma\Sigma^{\mathsf{T}}\widetilde{u}\|_1 \\
&= \|\gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\pi_{\widetilde{u}}} - \gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\widetilde{u}} + \gamma\Sigma^{\mathsf{T}}\widetilde{u}\|_1 \\
&\le \|\gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\pi_{\widetilde{u}}} - \gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\widetilde{u}}\|_1 + \|\gamma\Sigma^{\mathsf{T}}\widetilde{u}\|_1 \\
&= \|\gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}(\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}})\|_1 + \gamma\|\Sigma^{\mathsf{T}}\widetilde{u}\|_1 \\
&\le \|\gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}\|_1\|\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}}\|_1 + \gamma\|\Sigma^{\mathsf{T}}\widetilde{u}\|_1 \\
&= \gamma\|\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}}\|_1 + \gamma\|\Sigma^{\mathsf{T}}\widetilde{u}\|_1
\end{aligned}$$

Rearranging, we obtain $\|\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}}\|_1 \le \dfrac{\gamma\|\Sigma^{\mathsf{T}}\widetilde{u}\|_1}{(1-\gamma)}$. A similar calculation shows a bound in terms of $\Sigma^{\mathsf{T}}\mu_{\pi_{\widetilde{u}}}$:

$$\begin{aligned}
\|\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}}\|_1 &= \|(1-\gamma)\nu_0 + \gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\pi_{\widetilde{u}}} - [(1-\gamma)\nu_0 + \gamma P^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\widetilde{u}} - \gamma\Sigma^{\mathsf{T}}\widetilde{u}]\|_1 \\
&= \|(1-\gamma)\nu_0 + \gamma \widehat{P}^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\pi_{\widetilde{u}}} + \gamma\Sigma^{\mathsf{T}}\mu_{\pi_{\widetilde{u}}} - [(1-\gamma)\nu_0 + \gamma \widehat{P}^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\widetilde{u}}]\|_1 \\
&= \|(1-\gamma)\nu_0 + \gamma \widehat{P}^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\pi_{\widetilde{u}}} - (1-\gamma)\nu_0 - \gamma \widehat{P}^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\widetilde{u}} + \gamma\Sigma^{\mathsf{T}}\mu_{\pi_{\widetilde{u}}}\|_1 \\
&= \|\gamma \widehat{P}^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\pi_{\widetilde{u}}} - \gamma \widehat{P}^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\widetilde{u}} + \gamma\Sigma^{\mathsf{T}}\mu_{\pi_{\widetilde{u}}}\|_1 \\
&\le \|\gamma \widehat{P}^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\pi_{\widetilde{u}}} - \gamma \widehat{P}^{\mathsf{T}}_{\pi_{\widetilde{u}}}\nu_{\widetilde{u}}\|_1 + \|\gamma\Sigma^{\mathsf{T}}\mu_{\pi_{\widetilde{u}}}\|_1 \\
&= \|\gamma \widehat{P}^{\mathsf{T}}_{\pi_{\widetilde{u}}}(\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}})\|_1 + \gamma\|\Sigma^{\mathsf{T}}\mu_{\pi_{\widetilde{u}}}\|_1 \\
&\le \|\gamma \widehat{P}^{\mathsf{T}}_{\pi_{\widetilde{u}}}\|_1\|\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}}\|_1 + \gamma\|\Sigma^{\mathsf{T}}\mu_{\pi_{\widetilde{u}}}\|_1 \\
&= \gamma\|\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}}\|_1 + \gamma\|\Sigma^{\mathsf{T}}\mu_{\pi_{\widetilde{u}}}\|_1
\end{aligned}$$

Rearranging, we obtain $\|\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}}\|_1 \le \dfrac{\gamma\|\Sigma^{\mathsf{T}}\mu_{\pi_{\widetilde{u}}}\|_1}{(1-\gamma)}$. Combining the above bounds and recalling that $\|\widetilde{\mu}_{\pi_{\widetilde{u}}} - \widetilde{u}\|_1 = \|\nu_{\pi_{\widetilde{u}}} - \nu_{\widetilde{u}}\|_1$, we obtain

$$\|\mu_{\pi_{\widetilde{u}}} - \widetilde{u}\|_1 \le \frac{\gamma \min\{\|\Sigma^{\mathsf{T}}\widetilde{u}\|_1, \|\Sigma^{\mathsf{T}}\mu_{\pi_{\widetilde{u}}}\|_1\}}{(1-\gamma)},$$

thus completing the proof. ■

It is insightful to think of what happens when the optimal occupancy measure is feasible in the RALP, that is, when $\Sigma^{\mathsf{T}}u^* = 0$. Unfortunately, even in this case, Proposition 2 cannot guarantee to output a near-optimal policy. Supposing that $\widetilde{u}$ is an optimal solution of the RALP, the best guarantee on the performance of the extracted policy $\pi_{\widetilde{u}}$ that one can derive from the above result is the following:

$$\langle u^* - \mu_{\pi_{\widetilde{u}}}, r\rangle = \langle u^* - \widetilde{u}, r\rangle + \langle \widetilde{u} - \mu_{\pi_{\widetilde{u}}}, r\rangle \le \langle \widetilde{u} - \mu_{\pi_{\widetilde{u}}}, r\rangle \le \frac{\gamma\,\|r\|_\infty \min\{\|\Sigma^{\mathsf{T}}\widetilde{u}\|_1, \|\Sigma^{\mathsf{T}}\mu_{\pi_{\widetilde{u}}}\|_1\}}{(1-\gamma)}.$$

Here the first inequality follows from the optimality of $\widetilde{u}$ to the RALP objective and the feasibility of u^*, and the second inequality uses Proposition 2. As the right-hand side depends on quantities related to $\widetilde{u}$ that are not controlled by the RALP solution, it is unclear if these terms can ever be small enough to guarantee strong performance. We thus think of the above guarantee as a curious initial result that may serve as the foundation of further analyses concerning relaxed LP formulations for nearly-linear MDPs.

4.2 Nearly-realizable Q^*

We now return to the more general setting where only Assumptions 1 through 3 are supposed to hold.

First, we leverage on the definition of the optimal state-value function $V^*(x) = \max_a Q^*(x,a)$ to further reduce the number of constraints and consider an alternative objective to ensure proper comparison between Q-functions. Hence, we define $V_\theta(x) = \max_a Q_\theta(x,a)$ and introduce $\xi_0 \in \Delta_{\mathcal{Z}}$ an initial state-action distribution to obtain the following alternative RALP:

$$\begin{aligned}&\text{minimize}_\theta \quad (1-\gamma)\langle \xi_0\,, Q_\theta\rangle\\&\text{subject to} \quad \mathcal{U}Q_\theta \geq \mathcal{U}[r+\gamma PV_\theta] \qquad (\theta\in\mathbb{R}^d).\end{aligned}$$

Note that the definition of V_θ makes the constraint nonlinear. However, due to its obvious connection with the initial dual RALP and the fact that the nonlinear constraint can be written as the intersection of a finite number of linear constraints, we continue to refer to it as a linear program.

Let $\mathcal{M}_{\text{RALP}}$ denote the feasible set of the above LP. It is trivial to show that when Q^* is exactly realizable by the feature map as portrayed in Assumption 2 with $\varepsilon = 0$, Q^* is feasible and indeed optimal in the dual RALP provided ξ_0 is fully supported on all state-action pairs. Therefore, our main result is an approximation error bound on the RALP solution when $\varepsilon > 0$. Precisely, we suppose that there exists $\theta^* \in \mathbb{R}^d, \rho \in \mathbb{R}^Z$ such that $Q^* = Q_{\theta^*} + \rho$ and $\|\rho\|_\infty = \varepsilon$.

Let θ_{RALP} denote an optimal solution of the RALP and $Q_{\text{RALP}} = Q_{\theta_{\text{RALP}}}$. We prove that when Q^* is approximately realizable by the feature map, as portrayed in Proposition 5 below, the approximation error of the RALP solution is within a constant factor of the best approximation error to Q^*.

Proposition 5 *Let ξ_0 denote some initial state-action distribution. With $\mathbf{e}$ in the column span of Φ, $\gamma \in (0,1)$ and $\varepsilon = \inf_{\theta\in\mathbb{R}^d} \|Q^* - Q_\theta\|_\infty$ the approximation error, then:*

$$\|Q^* - Q_{RALP}\|_{1,\xi_0} \leq \frac{4\varepsilon}{(1-\gamma)^2}$$

We only provide a brief proof sketch of Proposition 5 below, and refer the reader to Appendix A.2 for more details.

Proof sketch This proof draws ideas from the proof technique of De Farias and Van Roy (2003) in Section 4.1 of their work. The main challenge is that, unlike in their proof, there is no clear relation between the feasible points of the RALP and Q^*. The main novelty in our analysis is addressing this challenge by introducing the operator $T' : \mathbb{R}^{XA} \to \mathbb{R}^{XA}$ defined as $T' = \mathcal{B}\mathcal{U}T^*_Q$, and observe that any θ satisfying the dual RALP constraints also satisfies

$$Q_\theta \geq T'Q_\theta.$$

This claim easily follows from left-multiplying the constraints of the RALP by $\mathcal{B}$ and observing that by Assumption 3, we have $\mathcal{B}\mathcal{U}\Phi = \Phi$. Now, by virtue of Assumption 3, we can easily see that the operator $\mathcal{B}\mathcal{U}$ is a nonexpansion in L_∞, and thus the composed operator $T' = \mathcal{B}\mathcal{U}T^*$ is a γ-contraction due to the contraction property of the Bellman optimality operator T^*. Furthermore, T' is a monotone operator due to the monotonicity of both T^* and $\mathcal{B}\mathcal{U}$ (which follows from it being a positive matrix). Now, since T' is a contraction, it has a unique fixed point Q' satisfying $T'Q' = Q'$, and we can easily show that all feasible points θ of the RALP satisfy $Q_\theta \geq Q'$. Indeed, we have

$$Q_\theta \geq T'Q_\theta \geq (T')^2 Q_\theta \geq (T')^\infty Q_\theta = Q'.$$

The proof now follows from adapting the techniques of De Farias and Van Roy (2003) to our setting. In particular, we show that there exists a feasible point of the RALP that is within $\mathcal{O}(\varepsilon)$ in L_∞-distance to Q^*. This feasible point is given as $\widetilde{\theta} = \theta^* + k\vartheta$ with $k = (1+\gamma)\varepsilon/(1-\gamma)$, and the associated Q-function clearly satisfies

$$\|Q^* - Q_{\widetilde{\theta}}\|_\infty \leq \frac{2\varepsilon}{(1-\gamma)}.$$

It then remains to show that $\widetilde{\theta}$ is not too far from the RALP solution Q_{RALP}. This can be done as

$$\|Q_{\widetilde{\theta}} - Q_{\text{RALP}}\|_{1,\xi_0} = \left\langle \xi_0, Q_{\widetilde{\theta}} - Q_{\text{RALP}}\right\rangle \leq \left\langle \xi_0, Q_{\widetilde{\theta}} - Q'\right\rangle = \|Q_{\widetilde{\theta}} - Q'\|_{1,\xi_0},$$

where the first step used the optimality of Q_{RALP}, and last two steps the fact that $Q_{\text{RALP}} \geq Q'$ and $Q_{\widetilde{\theta}} \geq Q'$ both hold due to feasibility of the two Q-functions. Overall, we have

$$\|Q^* - Q_{\text{RALP}}\|_{1,\xi_0} \leq \|Q^* - Q_{\widetilde{\theta}}\|_{1,\xi_0} + \|Q_{\widetilde{\theta}} - Q_{\text{RALP}}\|_{1,\xi_0} \leq \|Q^* - Q_{\widetilde{\theta}}\|_\infty + \|Q_{\widetilde{\theta}} - Q'\|_\infty$$

It then remains to bound the last term on the right-hand side. To this end, we first exploit the contraction property of T' to show that

$$\|Q_{\widetilde{\theta}} - Q'\|_\infty \le \frac{\|Q_{\widetilde{\theta}} - T' Q_{\widetilde{\theta}}\|_\infty}{(1-\gamma)}.$$

The term on the right-hand side can then be bounded by taking advantage of the properties of the operator $\mathcal{BU}$ as follows:

$$\|Q_{\widetilde{\theta}} - T' Q_{\widetilde{\theta}}\|_\infty = \|\mathcal{BU} Q_{\widetilde{\theta}} - \mathcal{BU} T^* Q_{\widetilde{\theta}}\|_\infty \le \|Q_{\widetilde{\theta}} - T^* Q_{\widetilde{\theta}}\|_\infty \le (1+\gamma)\|Q^* - Q_{\widetilde{\theta}}\|_\infty.$$

The first step here follows from $\mathcal{BU}\Phi = \Phi$ and the definition of T', the second step from the nonexpansion property of $\mathcal{BU}$, and the last step from straightforward algebraic manipulations involving the contraction property of T'. Using again our bound on $\|Q^* - Q_{\widetilde{\theta}}\|_\infty$, we thus obtain that

$$\|Q_{\widetilde{\theta}} - Q'\|_\infty \le \frac{2\varepsilon(1+\gamma)}{(1-\gamma)^2}, \tag{8}$$

which concludes the proof. ■

5. Related work

As we have seen already, like exact policy search and dynamic programming methods, the conventional LP formulation is known to suffer from the curse of dimensionality as the number of decision variables and constraints scale at least with the number of states. Hence, to disengage the statistical and computational complexities from the MDP size, substantial research is geared towards making functional modifications to the standard formulation while ensuring that the solution of the modified LP is within a reasonable distance to a true value function or better still the optimal value function. In this section, we review related works on two major modifications leading to Approximate Linear Programs (ALPs) and Reduced Approximate Linear Programs (RALPs).

From its inception, polynomial function approximation has paved way for empirical research beyond tabular methods. In the context of linear programming for optimal control, linear function approximation has been first introduced by Schweitzer and Seidmann (1985). As already pointed out in this early work, replacing the exact value representation in the dual objective with an approximate representation is known to impose a trade-off between computation cost and accuracy. Quantifying this accuracy trade-off for the LP setting, De Farias and Van Roy (2003) proved that in the discounted MDP setting, the approximate linear program proposed by Schweitzer and Seidmann returns a relatively good approximation to the optimal state-value function provided that the all-ones vector is within the column span of the feature matrix. In particular, the ALP is guaranteed to return the optimal state-value function if it is within the span of the features.

However, the number of constraints in the optimization problem is prohibitively large which prevents the development of truly efficient methods based on this LP. This issue has later been addressed using a variety of ideas like constraint generation (Schuurmans and Patrascu, 2001; Guestrin et al., 2003), constraint sampling (De Farias and Van Roy, 2004) and constraint selection (Lakshminarayanan et al., 2017; Shariff and Szepesvári, 2020), all resulting in Reduced Approximate Linear Programs (RALPs) with fewer constraints, feasibility and near-optimality guarantees. The works of Chen et al. (2018) and Bas-Serrano and Neu (2020) proposed a more general parametrization of the primal variables, derived guarantees on the quality of the optimal solutions, and also provided efficient algorithms for producing near-optimal solutions of the reduced approximate LP. These results are all proved under strong realizability conditions on the proposed function approximators parametrizing the primal and dual variables. We highlight here the work of Bas-Serrano and Neu (2020), who show that realizability of the optimal primal and dual variables by itself is not sufficient to derive near-optimal policies from solutions of a relaxed LP. This observation falls in line with the recent hardness results in the broader context of reinforcement learning under weak function approximation (Du et al., 2019; Weisz et al., 2021; Wang et al., 2021).

In this paper, we took specific interest in the constraint selection approach by Lakshminarayanan et al. (2017) and Shariff and Szepesvári (2020) imposed by a core state assumption. Under this geometric assumption jointly enforced by a low-rank feature matrix for estimating the value function (Lakshminarayanan et al., 2017) with the all-ones vector in its column span (Shariff and Szepesvári, 2020), both sets of authors discover that a RALP with constraints corresponding to states (and all actions) whose feature vectors are linearly independent in the feature matrix, has a bounded feasible region and with a sample efficient algorithm (Shariff and Szepesvári, 2020) yields a policy with value error worse than that of the ALP by only a constant factor of the best approximation to the optimal state value function.

6. Conclusion

We have studied a linear programming approach to global planning in large MDPs. In view of notable limitations of the conventional LP formulation of Markov Decision Processes (MDP)s we promote the LP formulation featuring Q-functions first proposed by Mehta and Meyn (2009) as an alternative for such tasks. Building on a variety of previous results on approximate linear programming for optimal control, we employed function approximation and a core state-action assumption to obtain a feasible Reduced Approximate Linear Program (RALP). Finally, we showed that the RALP even with fewer constraints can give a relatively good approximation to the optimal state-action value function. This result improves on previous work by showing that the LP approach is potentially suitable for computing globally near-optimal policies in large MDPs. An important next step is to develop an algorithm which for any γ-discounted MDP, produces an approximate solution of the corresponding RALP problem and returns a near-optimal policy using polynomial number of sample trajectories and computation time independent of the size of the state space.

Acknowledgments

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).

References

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

Joan Bas 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. PMLR, 2020.

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.

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.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored mdps. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

Arie Hordijk and LCM Kallenberg. Linear programming and markov decision chains. *Management Science*, 25(4): 352–362, 1979.

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. PMLR, 2020.

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.

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.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. pages 2315–2325, 2019a.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. AlgaeDICE: Policy gradient from arbitrary experience. *CoRR*, abs/1912.02074, 2019b.

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.

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

Dale Schuurmans and Relu Patrascu. Direct value-approximation for factored mdps. *Advances in Neural Information Processing Systems*, 14:1579–1586, 2001.

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.

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. PMLR, 2020.

Yuanhao Wang, Ruosong Wang, and Sham Kakade. An exponential lower bound for linearly realizable mdp with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 34, 2021.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

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.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. *Advances in neural information processing systems*, 26, 2013.

Appendix A. Appendix

A.1 The Approximate Linear Program

We recall here the Approximate Linear Program (ALP) studied by Neu and Pike-Burke (2020) and Bas-Serrano et al. (2021).

Under the Q^* realizability assumption, Assumption 2, it then becomes reasonable to optimize over the class of parametric functions of the form $Q_\theta = \Phi\theta$. Restricting the primal variables, u, μ to belong to the probability simplex $\Delta_{\mathcal{Z}}$. Plugging Q_θ into LP (3), we obtain the following ALP:

$$\begin{aligned} \max_{\mu,u} \quad & \langle \mu \,, r \rangle \\ \text{subject to} \quad & E^\mathsf{T} u = (1-\gamma)\nu_0 + \gamma P^\mathsf{T} \mu, \\ & \Phi^\mathsf{T}\mu = \Phi^\mathsf{T} u \,. \quad \mu, u \in \Delta_{\mathcal{Z}} \end{aligned}$$

With dual:

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

For the sake of our analysis of the primal RALP, we study the primal ALP. Let $\mathcal{M}$, $\mathcal{M}_\Phi$, $\widetilde{\mathcal{M}}_\Phi$ denote the feasible set of the primal LP (1), ALP and RALP respectively. Recall that $(\mu, u) = (u^*, u^*)$ is optimal in LP (1) and for any $(\mu, u) \in \mathcal{M}$, u is a valid occupancy measure satisfying Equation (2). Studying the feasible sets, $\mathcal{M}$ and $\mathcal{M}_\Phi$, it is easy to see that $\mathcal{M} \subseteq \mathcal{M}_\Phi$. Therefore $(u^*, u^*) \in \mathcal{M}_\Phi$. Introducing Assumption 4, one can then show that for any $(\mu, u) \in \mathcal{M}_\Phi$, u is a valid occupancy measure also satisfying Equation (2), and ultimately that (u^*, u^*) is optimal in the primal ALP (Neu and Pike-Burke, 2020; Bas-Serrano et al., 2021).

A.2 Proof of Proposition 5

To prove Proposition 5, we first show that there exists a feasible point of the RALP within $\mathcal{O}(\varepsilon)$ to Q^* in L_∞-distance. This is formally stated in the following lemma:

Lemma 6 *There exists $\theta \in \mathcal{M}_{RALP}$ such that*

$$\|Q^* - Q_\theta\|_\infty \leq \frac{2\varepsilon}{(1-\gamma)}$$

Proof The proof of Lemma 6 draws insights from section 4.1 of De Farias and Van Roy (2003).

First, by definition of the RALP feasible set, for all $\theta \in \mathcal{M}_{\text{RALP}}$,

$$\begin{aligned} & \mathcal{U}Q_\theta \geq \mathcal{U}[r + \gamma P V_\theta] \\ \implies \; & \mathcal{B}\mathcal{U}Q_\theta \geq \mathcal{B}\mathcal{U}[r + \gamma PMQ_\theta] \\ \implies \; & \quad Q_\theta \geq \mathcal{B}\mathcal{U}[r + \gamma PMQ_\theta] = \mathcal{B}\mathcal{U}T^*_Q Q_\theta \end{aligned} \tag{9}$$

Set $T' = \mathcal{B}\mathcal{U}T^*_Q$. Clearly, $T' : \mathbb{R}^Z \to \mathbb{R}^Z$ is a $\|\cdot\|_\infty$-contraction mapping of factor γ as $\mathcal{B}$ and $\mathcal{U}$ are stochastic matrices and by the contraction property of T^*_Q.

Recall that by the realizability assumption Assumption 2:

$$\|Q^* - Q_{\theta^*}\|_\infty = \varepsilon \implies |Q^* - Q_{\theta^*}| \leq \varepsilon \mathbf{e} \implies Q^* \leq Q_{\theta^*} + \varepsilon \mathbf{e}$$

Since $Q^* = T^*_Q Q^*$, also by monotonicity of $\mathcal{B}\mathcal{U}$ and the property and $\mathcal{B}\mathcal{U}\mathbf{e} = \mathbf{e}$

$$\begin{aligned} T'Q^* &\leq \mathcal{B}\mathcal{U}Q_{\theta^*} + \varepsilon\mathcal{B}\mathcal{U}\mathbf{e} \\ &= Q_{\theta^*} + \varepsilon\mathbf{e} \\ \implies T'Q^* &\leq Q_{\theta^*} + \varepsilon\mathbf{e} \end{aligned} \tag{10}$$

Again, by the contraction property of the Bellman optimality operator T_Q^*:

$$\|T_Q^* Q_{\theta^*} - Q^*\|_\infty \leq \gamma\varepsilon \implies |T_Q^* Q_{\theta^*} - Q^*| \leq \gamma\varepsilon\mathbf{e} \implies T_Q^* Q_{\theta^*} \leq Q^* + \gamma\varepsilon\mathbf{e}$$

By monotonicity of $\mathcal{BU}$ and the property $\mathcal{BU}\mathbf{e} = \mathbf{e}$

$$T' Q_{\theta^*} \leq T' Q^* + \gamma\varepsilon\mathbf{e} \tag{11}$$

For any $g \in \mathbb{R}^Z$, $k \in \mathbb{R}$, by definition of T' in Equation (9),

$$\begin{aligned} T'(g + k\mathbf{e}) &= \mathcal{BU}[r + \gamma PM(g + k\mathbf{e})] \\ &= \mathcal{BU}[r + \gamma PMg + \gamma kP\mathbf{e}] \\ &= \mathcal{BU}[r + \gamma PMg] + \gamma k\mathcal{BU}\mathbf{e} \\ &= T'g + \gamma k\mathbf{e} \end{aligned}$$

$$T'(g + k\mathbf{e}) = T'g + \gamma k\mathbf{e} \tag{12}$$

Then,

$$\begin{aligned} T'(Q_{\theta^*} + k\mathbf{e}) &= T'Q_{\theta^*} + \gamma k\mathbf{e} \\ &\leq T'Q^* + \gamma\varepsilon\mathbf{e} + \gamma k\mathbf{e} && \text{By Equation (11)} \\ &\leq Q_{\theta^*} + \varepsilon\mathbf{e} + \gamma\varepsilon\mathbf{e} + \gamma k\mathbf{e} && \text{By Equation (10)} \\ &\leq Q_{\theta^*} + (1+\gamma)\varepsilon\mathbf{e} + \gamma k\mathbf{e} \\ &\leq Q_{\theta^*} + k\mathbf{e} + (1+\gamma)\varepsilon\mathbf{e} + \gamma k\mathbf{e} - k\mathbf{e} \\ &\leq Q_{\theta^*} + k\mathbf{e} + (1+\gamma)\varepsilon\mathbf{e} - (1-\gamma)k\mathbf{e} \end{aligned}$$

Let $\widetilde{\theta} = \theta^* + k\vartheta$. By Assumption 1, $Q_{\widetilde{\theta}} = Q_{\theta^*} + k\mathbf{e}$. Then, for $k = \dfrac{(1+\gamma)\varepsilon}{(1-\gamma)}$, $\widetilde{\theta} \in \mathcal{M}_{\text{RALP}}$ since $Q_{\widetilde{\theta}} \geq T'Q_{\widetilde{\theta}}$, satisfying the sole feasibility condition. Since, by definition of $\widetilde{\theta}$ and choice of k we have that:

$$\begin{aligned} Q_{\widetilde{\theta}} &= Q_{\theta^*} + \frac{(1+\gamma)\varepsilon\mathbf{e}}{(1-\gamma)} \\ &\geq Q^* - \varepsilon\mathbf{e} + \frac{(1+\gamma)\varepsilon\mathbf{e}}{(1-\gamma)} && \text{By definition of } Q^* \\ &= Q^* + \frac{2\gamma\varepsilon\mathbf{e}}{(1-\gamma)} \end{aligned}$$

Since the second term is positive, $Q_{\widetilde{\theta}} \geq Q^*$ and we can further bound the error in absolute terms as follows:

$$\begin{aligned} \|Q_{\widetilde{\theta}} - Q^*\|_\infty &= \|Q_{\widetilde{\theta}} - Q_{\theta^*} + Q_{\theta^*} - Q^*\|_\infty \\ &\leq \|Q_{\widetilde{\theta}} - Q_{\theta^*}\|_\infty + \|Q_{\theta^*} - Q^*\|_\infty \\ &= \left\| \frac{(1+\gamma)\varepsilon}{(1-\gamma)}\mathbf{e} \right\|_\infty + \varepsilon \\ &= \frac{(1+\gamma)\varepsilon}{(1-\gamma)} + \varepsilon \\ &= \frac{(1+\gamma)\varepsilon + (1-\gamma)\varepsilon}{(1-\gamma)} \\ &= \frac{2\varepsilon}{(1-\gamma)} \end{aligned}$$

Therefore, for some $\theta \in \mathcal{M}_{\text{RALP}}$, $\|Q_{\widetilde{\theta}} - Q^*\|_\infty \leq \dfrac{2\varepsilon}{(1-\gamma)}$. This concludes the proof. ■

Now, to prove Proposition 5:

Proof In proving Lemma 6, we introduced an operator $T' : \mathbb{R}^Z \to \mathbb{R}^Z$ with $T' = \mathcal{BU}T_Q^*$ acting on vectors $g \in \mathbb{R}^Z$ such that $T'g = \mathcal{BU}[r + \gamma PMg]$. By the monotonicity of $\mathcal{BU}$ and the contraction property of T_Q^*, it is easy to show that T' is also a contraction map with constant factor γ. So, it has a unique fixed point. Let Q' denote the fixed point of T'. That is, $T'Q' = Q'$.

Since the feasibility condition of the RALP can be rewritten to imply that for all $\theta \in \mathcal{M}_{\text{RALP}}$, $Q_\theta \geq T'Q_\theta$, we know that $Q_\theta \geq Q'$ for all $\theta \in \mathcal{M}_{\text{RALP}}$. With reference to $\widetilde{\theta} \in \mathcal{M}_{\text{RALP}}$ for which Lemma 6 is satisfied, our first line of action is to bound the maximum error between $Q_{\widetilde{\theta}}$ and the fixed point of T' with the following steps:

$$\begin{aligned}
\|Q_{\widetilde{\theta}} - Q'\|_\infty &= \|Q_{\widetilde{\theta}} - T'Q_{\widetilde{\theta}} + T'Q_{\widetilde{\theta}} - Q'\|_\infty \\
&\leq \|Q_{\widetilde{\theta}} - T'Q_{\widetilde{\theta}}\|_\infty + \|T'Q_{\widetilde{\theta}} - Q'\|_\infty \\
&\leq \|Q_{\widetilde{\theta}} - T'Q_{\widetilde{\theta}}\|_\infty + \gamma\|Q_{\widetilde{\theta}} - Q'\|_\infty
\end{aligned}$$

By rearranging, we obtain

$$\|Q_{\widetilde{\theta}} - Q'\|_\infty \leq \frac{\|Q_{\widetilde{\theta}} - T'Q_{\widetilde{\theta}}\|_\infty}{(1-\gamma)}. \tag{13}$$

We can then bound the term appearing on the right-hand side by taking advantage of the properties of the operator $\mathcal{BU}$ as follows:

$$\begin{aligned}
\|Q_{\widetilde{\theta}} - T'Q_{\widetilde{\theta}}\|_\infty &= \|\mathcal{BU}Q_{\widetilde{\theta}} - \mathcal{BU}T_Q^*Q_{\widetilde{\theta}}\|_\infty \\
&\leq \|\mathcal{BU}\|_\infty \|Q_{\widetilde{\theta}} - T_Q^*Q_{\widetilde{\theta}}\|_\infty \\
&= \|Q_{\widetilde{\theta}} - T_Q^*Q_{\widetilde{\theta}}\|_\infty \\
&= \|Q_{\widetilde{\theta}} - Q^* + Q^* - T_Q^*Q_{\widetilde{\theta}}\|_\infty \\
&\leq \|Q_{\widetilde{\theta}} - Q^*\|_\infty + \|Q^* - T_Q^*Q_{\widetilde{\theta}}\|_\infty \\
&\leq \|Q_{\widetilde{\theta}} - Q^*\|_\infty + \gamma\|Q^* - Q_{\widetilde{\theta}}\|_\infty \\
&= (1+\gamma)\|Q^* - Q_{\widetilde{\theta}}\|_\infty \\
&\leq \frac{2\varepsilon(1+\gamma)}{(1-\gamma)}.
\end{aligned}$$

The last inequality follows from Lemma 6. Substituting this in Equation 13, we have that:

$$\|Q_{\widetilde{\theta}} - Q'\|_\infty \leq \frac{2\varepsilon(1+\gamma)}{(1-\gamma)^2} \tag{14}$$

Now we have established a relationship between $Q_{\widetilde{\theta}}$ and Q', bearing in mind that $Q_{\text{RALP}} \geq Q'$ since $\theta_{\text{RALP}} \in \mathcal{M}_{\text{RALP}}$ and Q' is a unique fixed point of T', and $Q_{\widetilde{\theta}} \geq Q^*$, a side result from Lemma 6, we proceed to prove the main result:

$$\begin{aligned}
\|Q^* - Q_{\text{RALP}}\|_{1,\xi_0} &= \|Q^* - Q_{\widetilde{\theta}} + Q_{\widetilde{\theta}} - Q_{\text{RALP}}\|_{1,\xi_0} \\
&\leq \|Q^* - Q_{\widetilde{\theta}}\|_{1,\xi_0} + \|Q_{\widetilde{\theta}} - Q_{\text{RALP}}\|_{1,\xi_0} \\
&= \|Q^* - Q_{\widetilde{\theta}}\|_\infty + \langle \xi_0 \,, |Q_{\widetilde{\theta}} - Q_{\text{RALP}}| \rangle \\
&\leq \frac{2\varepsilon}{(1-\gamma)} + \langle \xi_0 \,, Q_{\widetilde{\theta}} - Q_{\text{RALP}} \rangle \\
&\leq \frac{2\varepsilon}{(1-\gamma)} + \langle \xi_0 \,, Q_{\widetilde{\theta}} - Q' \rangle \\
&= \frac{2\varepsilon}{(1-\gamma)} + \langle \xi_0 \,, |Q_{\widetilde{\theta}} - Q'| \rangle \\
&\leq \frac{2\varepsilon}{(1-\gamma)} + \|Q_{\widetilde{\theta}} - Q'\|_\infty \\
&\leq \frac{2\varepsilon}{(1-\gamma)} + \frac{2\varepsilon(1+\gamma)}{(1-\gamma)^2} \\
&= \frac{4\varepsilon}{(1-\gamma)^2}
\end{aligned}$$

This completes the proof. ■