A Linearly Relaxed Approximate Linear Program for Markov Decision Processes

Chandrashekar Lakshminarayanan, Shalabh Bhatnagar, and Csaba Szepesvári

***Abstract*—Approximate linear programming (ALP) and its variants have been widely applied to Markov decision processes (MDPs) with a large number of states. A serious limitation of ALP is that it has an intractable number of constraints, as a result of which constraint approximations are of interest. In this paper, we define a linearly relaxed approximation linear program (LRALP) that has a tractable number of constraints, obtained as positive linear combinations of the original constraints of the ALP. The main contribution is a novel performance bound for LRALP.**

***Index Terms*—Approximate linear programming (ALP), Markov decision processes (MDPs).**

I. Introduction

Markov decision processes (MDPs) have proved to be an indispensable model for sequential decision making under uncertainty with applications in networking, traffic control, robotics, operations research, business, finance, artificial intelligence, health-care, and more (see, e.g., [1]–[10]). In this paper, we adopt the framework of discrete-time, discounted MDPs when a controller steers the stochastically evolving state of a system while receiving rewards that depend on the states visited and actions chosen. The goal is to choose the actions so as to maximize the *return*, defined as the total discounted expected reward. A controller that uses past state information is called a *policy*. An *optimal policy* is one that maximizes the value, no matter where the process is started from [7]. In this paper, we consider planning problems where the goal is to calculate actions of policies that give rise to high values and give new error bounds on the quality of solutions obtained by solving linear programs of tractable size. To explain the contributions in more detail, we start by describing the computational challenges involved in planning.

The main objective of *planning* is to compute actions of an optimal policy while interacting with an MDP model. In finite state-action MDPs, assuming access to individual transition probabilities and rewards along transitions, various algorithms are available to perform this computation in time and space that scales polynomially with the number of states and actions. However, in most practical applications, the MDP is compactly represented and if it is not infinite, the number of states scale *exponentially* with the *size of the representation* of the MDP. If planners are allowed to perform some fixed number of calculations for each state encountered, it is possible to use sampling to make the per-state calculation-cost independent of the size of the state space [11]–[13]. Nevertheless, the resulting methods are still quite limited. In fact, various hardness results show that computing actions of (near-) optimal policies is intractable in various senses and in various compactly represented MDPs [14]. Given these negative results, it is customary to adopt the *modest goal of efficiently computing actions of a policy that is nearly as good as a policy chosen by a suitable (computationally unbounded, and well-informed) oracle from a given restricted policy class*. Here, within some restrictions (see below), the policy class can be chosen by the user. The more flexibility the user is given in this choice, the stronger a planning method is.

A popular approach along these lines, which goes back to Schweitzer and Seidmann [15], relies on considering linear approximations to the *optimal value function*: The idea is that, similarly to linear regression, a fixed sequence of basis functions are combined linearly. The user's task is to use *a priori* knowledge of the MDP to choose the basis functions so that a good approximation to the optimal value function will exist in the linear space spanned by the basis functions. The idea then is to design some algorithm to find the coefficients of the basis functions that give a good approximation, while keeping computation cost in check. Finding a good approximation is sufficient, since at the expense of an extra $O(1/\varepsilon^2)$ randomized computation, a uniform $O(\varepsilon)$-approximation to the optimal value function can be used to calculate an action of an $O(\varepsilon)$-optimal policy at any given state (e.g., follow the ideas in [12] and [13]; see also [16, Theorem 3.7]). Since the number of coefficients can be much smaller than the number of states, the algorithms that search for the coefficients have the potential to run efficiently regardless of the number of states.

Following Schweitzer and Seidmann [15], most of the literature considers algorithms that are obtained from restricting exact planning methods to search in the span of the fixed basis functions when performing computations. In this paper, we consider the so-called *approximate linear programming* (ALP) approach, which was heavily studied during the last two decades, e.g., [17]–[27]. The basic idea here is to combine a linear program whose solution is the optimal value function (and thus the number of optimization variables in it scales with the number of states) with a linear constraint that restricts the optimization variables to lie in the subspace spanned by the basis functions. As already noted by Schweitzer and Seidmann [15], the new LP can still be kept feasible by just adding one special basis function, while by substituting the "value function candidates" with their linear expansions, the number of optimization variables becomes the number of basis functions. As shown by de Farias and Van Roy [19], the solution to the resulting LP is within a constant factor of the best approximation to the optimal

Manuscript received March 24, 2017; revised August 3, 2017; accepted August 13, 2017. Date of publication August 22, 2017; date of current version March 27, 2018. Recommended by Associate Editor L. H. Lee. *(Corresponding author: Chandrashekar Lakshminarayanan.)*

C. Lakshminarayanan and C. Szepesvári are with the Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada (e-mail: cnarayan@ualberta.ca; csaba.szepesvari@ualberta.ca).

S. Bhatnagar is with the Department of Computer Science and Automation and Robert Bosch Centre for Cyber Physical Systems, Indian Institute of Science, Bengaluru 560012, India (e-mail: shalabh@iisc.ac.in).

Digital Object Identifier 10.1109/TAC.2017.2743163

value function within the span of the chosen bases. However, since the number of constraints in the LP is still proportional to the number of states, it is not obvious whether a solution to the resulting LP can be found in time independent of the number of states (other computations can be done in time independent of the number of states, e.g., using sampling, at the price of a controlled increase in the error, see e.g., [22, Theorem 6]).

Most of the literature is thus devoted to designing methods to select a tractable subset of the constraints while keeping the approximation guarantees, as well as keeping computations tractable. Since a linear objective is optimized by a point on the boundary of the feasible region, knowing the optimizer would be sufficient to eliminate all but as many constraints as the number of optimization variables. The question is how to find a superset of these, or an approximating set, without incurring much computational overhead. Schuurmans and Patrascu [17] and Guestrin *et al.* [18] propose constraint generation in a setting where the MDP has additional structure (i.e., factorized transition structure). This additional structure is then exploited in designing constraint generation methods which are able to efficiently generate violated constraints. A more general approach due to de Farias and Van Roy [20] is to choose a random subset of the constraints by choosing states to be included at random from a distribution that reflects the "importance" of states. While constraint generation can be powerful, it is not known how solution quality degrades with the budget on the constraints generated (Guestrin *et al.* [18] note that the number of constraints generated can be at most exponential in a fundamental quantity, the induced width of a so-called cost-network, which may be large and is in general hard to control). For constraint sampling, de Farias and Van Roy [20] prove a bound on the suboptimality, but this bound applies only in the unrealistic scenario when the constraints are sampled from an *idealized* distribution, which is related to the stationary distribution of an optimal policy. While it is possible to extend this result to any sampling distribution, the bound scales with the mismatch between the sampling and the idealized distributions, which, in general, will be uncontrolled. Another weakness of the bound is that when constraints are dropped, the linear program may become unbounded. To prevent this, de Farias and Van Roy [20] propose imposing an extra constraint on the optimization variables. The bound they obtain, however, scales with the *worst approximation error* over this constraint set. While in a specific example it is shown that this error can be controlled, no general results are derived in this direction. Later works, such as of Desai *et al.* [23] and Bhat *et al.* [26], repeat the analysis of de Farias and Van Roy [20] in combinations with other ideas. However, no existing work that we know of addresses the above weaknesses of the result of de Farias and Van Roy [20].

Our main contribution is a new suboptimality bound for the case when the constraint system is replaced with a smaller, linearly projected constraint system. We also propose a specific way of adding the extra constraint to keep the resulting LP bounded. Rather than relying on combinatorial arguments (such as those at the heart of de Farias and Van Roy [20]), our argument uses previously unexploited geometric structure of the linear programs underlying MDPs. As a result, our bound avoids distribution-mismatch terms and we also remove the scaling with worst approximation error. A specific outcome of our general result is the realization that it is beneficial to select states so that the "feature vectors" of all states when scaled with a fixed constant factor are included in the conic hull of the "feature vectors" underlying the selected states. This suggests to choose the basis functions so that this property can be satisfied by selecting only a few states. As we will argue, this property holds for several popular choices of basis functions. A preliminary version of this paper without the theoretical analysis and without the geometric arguments was published in a short conference communication [28].

II. Background

The purpose of this section is to introduce the necessary background before we can present the problem studied and the main results.

We shall consider finite state-action space, discounted total expected reward MDPs. We note in passing that the assumption that number of states is finite is mainly made for convenience and at the expense of a more technical presentation could be lifted. We will comment later on the assumption concerning the number of actions. Let the set of states, or state space be $\mathcal{S} = \{1, 2, \dots, S\}$ and let the set of actions be $\mathcal{A} = \{1, 2, \dots, A\}$. For simplicity, we assume that all actions are admissible in all states. Given a choice of an action $a \in \mathcal{A}$ in a state $s \in \mathcal{S}$, the controller incurs a reward (or gain) of $g_a(s) \in [0, 1]$ and the state moves to a next state $s' \in \mathcal{S}$ with probability $p_a(s, s')$. A *policy* u is a mapping from states to actions.[1] When a policy is followed, the state sequence evolves as a Markov chain with transition probabilities given by P_u matrix whose (s, s')th entry is $P_{u(s)}(s, s')$. Along the way the rewards are generated from g_u defined by $g_u(s) \doteq g_{u(s)}(s)$. The *value* of following a policy from a starting state s is denoted by $J_u(s)$ and is defined as the expected total discounted reward. Thus

$$J_u(s) \doteq \sum_{t=0}^{\infty} \alpha^t (P_u^t g_u)(s)$$

where $\alpha \in (0, 1)$ is the so-called discount factor. We call J_u the *value function* of policy u. The value function of a policy satisfies the fixed-point equation $J_u = T_u J_u$ where the affine-linear operator T_u is defined by $T_u J = g_u + \alpha P_u J$. An *optimal policy*, is one that maximizes the value simultaneously for all initial states. The *optimal value function* J^* is defined by $J^*(s) = \max_u J_u(s)$ and is known to be the solution of the fixed-point equation $J^* = T J^*$ where the operator T is defined by $(T J)(s) = \max_u (T_u J)(s)$, $s \in \mathcal{S}$, i.e., the maximization is componentwise. Optimal policies exist and in fact any policy u such that the equation $T_u J^* = T J^*$ holds is optimal (e.g., [16, Corollary 3.3]). A policy u is said to be *greedy* with respect to (w.r.t.) J if $T_u J = T J$. Thus, any policy that is greedy w.r.t. J^* is optimal.

III. Linearly Relaxed ALP

In this section, we introduce the computational model used and the "Linearly relaxed approximate linear program" a relaxation of the ALP.

As discussed in the introduction, we are interested in methods that compute a good approximation to the optimal value function. As noted earlier, at the expense of a modest additional cost, knowing an $O(\varepsilon)$ approximation to J^* at a few states suffices to compute actions of an $O(\varepsilon)$-optimal policy. We will take a more general view, and we will consider calculating good approximations to J^* with respect to a weighted 1-norm, where the weights c form a probability distribution over $\mathcal{S}$. Recall that the weighted 1-norm $\|J\|_{1,c}$ of a vector $J \in \mathbb{R}^S$ is defined as $\|J\|_{1,c} = \sum_s c(s)|J(s)|$. Note that here and in what follows, we identify elements of $\mathbb{R}^{\mathcal{S}}$ (functions, mapping $\mathcal{S} = \{1, \dots, S\}$ to the reals) with elements of $\mathbb{R}^S$ in the obvious way. This allows us to write, e.g., $c^\top J$, which denotes $\sum_s c(s) J(s)$.

To introduce the optimization problem we study, first recall that the optimal value function J^* is the solution of the fixed point equation $T J^* = J^*$. It follows from the definition of T that $J^* = \max_u T_u J^* \geq T_u J^*$ for any u, where $\geq$ is the componentwise partial ordering of vectors ($\leq$ is the reverse relation). With some abuse of notation, we also introduce T_a to denote T_u where $u(s) = a$ for any $s \in \mathcal{S}$. It follows

[1] For the scope of this paper, it suffices to restrict our attention to such policies as opposed to considering history dependent policies. See Chapter 3, and specifically [16, Corollary 3.3].

that $J^* \geq T_a J^*$ for any $a \in \mathcal{A}$ and also that $T = \max_a T_a$, where again the maximization is componentwise. We call a vector J that satisfies $J \geq T_a J$ for any $a \in A$ *superharmonic*. Note that this is a set of linear inequalities. By our note on T and $(T_a)_a$, these inequalities can also be written compactly as $J \geq TJ$. It is not hard to show then that J^* is the smallest superharmonic function (i.e., for any J superharmonic, $J \geq J^*$). It also follows that for any $c \in \mathbb{R}^S_{++} \doteq (0, \infty)^S$, the unique solution to the linear program $\min\{c^\top J : J \geq TJ\} = \min\{c^\top J : J \geq T_a J, a \in \mathcal{A}\}$ is J^*.

Now, let $\phi_1, \ldots, \phi_k : \mathcal{S} \to \mathbb{R}$ be k basis functions. The ALP of Schweitzer and Seidmann [15] is obtained by adding the linear constraints $J = \sum_{i=1}^k r_i \phi_i$ to the above linear program. Eliminating J then gives $\min\{\sum_i r_i c^\top \phi_i : \sum_i r_i \phi_i \geq g_a + \alpha \sum_i r_i P_a \phi_i, a \in \mathcal{A}, r = (r_i) \in \mathbb{R}^k\}$. As noted by Schweitzer and Seidmann [15], the linear program is feasible as long as $\mathbf{1}$, defined as the vector with all components being identically equal to one, is in the span of $\{\phi_1, \ldots, \phi_k\}$. *For the purpose of computations, it is assumed that the values $c^\top \phi_i$, $i = 1, \ldots, k$ and the values $(P_a \phi_i)(s)$ and $g_a(s)$ can be accessed in constant time.* This assumption can be relaxed to assuming that one can access $g_a(s)$ and $\phi_i(s)$ for any (s, a) in constant time, as well as to that one can efficiently sample from c, from $P_a(s, \cdot)$ for any (s, a) pair, but the details of this are beyond the scope of the present work. As shown by de Farias and Van Roy [19], if r_{ALP} denotes the solution to the above ALP, then for $J_{\mathrm{ALP}} \doteq \sum_i r_{\mathrm{ALP}}(i)\phi_i \doteq \Phi r_{\mathrm{ALP}}$ it holds that $\|J_{\mathrm{ALP}} - J^*\|_{1,c} \leq \frac{2\varepsilon}{1-\alpha}$ provided $c^\top \mathbf{1} = 1$ and where $\varepsilon = \inf_r \|J^* - \Phi r\|_\infty$ is the error of approximating the optimal value with the span of the basis functions $\phi_1, \ldots, \phi_k$ and $\|J\|_\infty = \max_s |J(s)|$ is the maximum norm and $\Phi \in \mathbb{R}^{S \times k}$ is the matrix formed by $(\phi_1, \ldots, \phi_k)$. That the error of approximating J^* with J_{ALP} is $O(\varepsilon)$ is significant: The user can focus on finding a good basis, leaving the search for the "right" coefficients to a linear program solver.

While solving the ALP can be significantly cheaper than solving the LP underlying the MDP and thus it can be advantageous for moderate-scale MDPs, the number of constraints in the ALP is SA, hence the ALP is still intractable for huge-scale MDPs. To reduce the number of constraints, we consider a relaxation of ALP where the constraints are replaced with positive linear combinations of them. Recalling that the constraints took the form $J \geq g_a + \alpha P_a J$ (with $J = \Phi r$), choosing m to be target number of constraints, for $1 \leq i \leq m$, the ith new constraint is given by $\sum_a w_{i,a}^\top J \geq \sum_a w_{i,a}^\top (g_a + \alpha P_a J)$, where the choice of m and that of the vectors $w_{i,a} \in \mathbb{R}^S_+$ is left to the user. Note that this results in a linear program with k variables and m constraints, which can be written as

$$\begin{aligned} &\min_{r \in \mathbb{R}^k} \; c^\top \Phi r \\ &\text{s.t.} \sum_a W_a^\top \Phi r \geq \sum_a W_a^\top (g_a + \alpha P_a \Phi r) \end{aligned} \tag{1}$$

where $W_a = (w_{1,a}, \ldots, w_{m,a}) \in \mathbb{R}_+^{S \times m}$. Note that the (i, j)th entry of the $m \times k$ constraint matrix of the resulting LP is $\sum_a w_{i,a}^\top \phi_j - \alpha \sum_a w_{i,a}^\top P_a \phi_j$ and assuming that $(w_{i,a})_a$ has p nonzero elements, this can be calculated in $O(p)$ time, making the total cost of obtaining the constraint matrix to be $O(mkp)$ regardless of the values of S and A.

We will call the LP in (1) the *linearly relaxed approximate linear program (LRALP)*. Any LP obtained using any constraint selection/generation process can be represented by choosing an appropriate binary-valued matrix $W^\top = (W_1^\top, \ldots, W_A^\top) \in \mathbb{R}_+^{m \times SA}$. In particular, when the constraints are selected in a random process as suggested by de Farias and Van Roy [20], the matrix W would be a random, binary-valued matrix.

Note that the LRALP may be unbounded. Unboundedness could be avoided by adding an extra constraint of the form $r \in \mathcal{N}$ to the LRALP, for a properly chosen polyhedron $\mathcal{N} \subset \mathbb{R}^k$.[2] However, it seems to us that it is downright misleading to think that guaranteeing a bounded solution will also lead to reasonable solutions. Thus we will stick to the above simple form, forcing a discussion of how W should be chosen to get meaningful results.[3]

Further insight into the choice of W can be gained by considering the Lagrangians of the ALP and LRALP. To write both LP's in a similar form let us introduce $E = (I_{S \times S}, \ldots, I_{S \times S})^\top$, where $I_{S \times S}$ is the $S \times S$ identity matrix. Further, let $H : \mathbb{R}^S \to \mathbb{R}^{SA}$ be the operator defined by $(HJ)^\top = ((T_1 J)^\top, \ldots, (T_A J)^\top)$. Note that H is an affine linear operator, which we call the *linear Bellman operator*. Then, the ALP can be written as $\min\{c^\top \Phi r \mid E\Phi r \geq H\Phi r\}$, while LRALP takes the form $\min\{c^\top \Phi r \mid W^\top E\Phi r \geq W^\top H\Phi r\}$. Hence, their Lagrangians are $\mathcal{L}_{\mathrm{ALP}}(r, \lambda) = c^\top \Phi r + \lambda^\top (H\Phi r - E\Phi r)$ and $\mathcal{L}_{\mathrm{LRALP}}(r, q) = c^\top \Phi r + q^\top W^\top (H\Phi r - E\Phi r)$, respectively. Thus, we can view Wq as a "linear approximation" to the dual variable $\lambda \in \mathbb{R}_+^{SA}$. This suggests that perhaps W should be chosen such that it approximates well the optimal dual variable. If Φ spans $\mathbb{R}^S$, the optimal dual variable λ^* is known to be the discounted occupancy measure underlying the optimal policy ([16, Theorem 3.18]), suggesting that the role of W is very similar to the role of Φ excepts that the subspace spanned by the columns of W should ideally be close to λ^*.

IV. Main Results

The purpose of this section is to present our main results. Let r_{LRA} be a solution to the LRALP given by (1) and let $J_{\mathrm{LRA}} = \Phi r_{\mathrm{LRA}}$. When multiple solutions exist, we can choose any of them. For the result, we assume that the LRALP is not unbounded, and hence a solution exists. In fact, we will assume something much stronger. The discussion of why our assumptions are reasonable and how to ensure that they hold is postponed to after the presentation of our results. Our main results bound the error $\|J^* - J_{\mathrm{LRA}}\|_{1,c}$.

The bound is given in terms of the approximation error of J^* with the basis functions $\Phi = (\phi_1, \ldots, \phi_k)$, as well as the deviation between two functions, $J^*_{\mathrm{ALP}}, J^*_{\mathrm{LRA}} : \mathcal{S} \to \mathbb{R}$, which we define next. In particular

$$J^*_{\mathrm{ALP}}(s) = \min\{r^\top \phi(s) \mid \Phi r \geq J^*, \; r \in \mathbb{R}^k\},$$

$$J^*_{\mathrm{LRA}}(s) = \min\{r^\top \phi(s) \mid W^\top E\Phi r \geq W^\top E J^*, \; r \in \mathbb{R}^k\}$$

where $s \in \mathcal{S}$. Recall that $E : \mathbb{R}^S \to \mathbb{R}^{SA}$ is defined so that $(EJ)^\top = (J^\top, \ldots, J^\top)$, i.e., E stacks its argument A-fold. Hence, $W^\top E = \sum_a W_a^\top$. Our strong assumption is that J^*_{LRA} is finite-valued. Note that $J^*_{\mathrm{ALP}} \geq J^*$ reflects the error due to using the basis functions $(\phi_j)_j$, and the magnitude of the deviation $J^*_{\mathrm{LRA}} - J^*_{\mathrm{ALP}}$ reflects the error introduced due to the relaxed constraint system.

Following [19] and [20], in the result the magnitude of the error $J^*_{\mathrm{LRA}} - J^*_{\mathrm{ALP}}$ and also that of the error of approximating J^* with the subspace spanned by Φ, will be measured in terms of a *weighted maximum norm*, $\|J\|_{\infty,\psi} = \max_{s \in \mathcal{S}} |J(s)|/\psi(s)$, where $\psi : \mathcal{S} \to \mathbb{R}_{++}$

[2] In particular, to obtain their theoretical result, de Farias and Van Roy [20] need the assumption that the set $\mathcal{N}$ is bounded and that it contains r_{ALP}. In fact, the error bound derived by de Farias and Van Roy [20] depends on the *worst* error of approximating J^* with Φr when r ranges over $\mathcal{N}$. Hence, if $\mathcal{N}$ is unbounded, their bound is vacuous. In the context of a particular application, de Farias and Van Roy [20] demonstrate that $\mathcal{N}$ can be chosen properly to control this term. However, no general construction is presented to choose $\mathcal{N}$.

[3] The only question is whether there is some value in adding constraints beyond choosing W properly. Our position is that the set $\mathcal{N}$ would most likely be chosen based on very little and general information; the useful knowledge is in choosing W, not in choosing some general set $\mathcal{N}$. Since randomization does not guarantee bounded solutions, de Farias and Van Roy [19] must use $\mathcal{N}$: In their case, $\mathcal{N}$ incorporates all the knowledge that makes the LP bounded.

is a positive-valued weighting function.[4] As also stressed by de Farias and Van Roy [19], the appropriate choice of ψ is crucial for MDPs with huge state-spaces: The problem is that if the range of values of $|J^*(s)|$ in different parts of the state space differ in orders of magnitude, it is not meaningful to expect to control the error of approximating it uniformly. By choosing the weighting function to reflect the magnitude of J^*, controlling the weighted maximum norm is achieved by controlling relative errors, which may be much easier to ensure than controlling meaningfully small absolute errors.

Just like de Farias and Van Roy [19], we will also require that ψ is a *stochastic Lyapunov-function* for the MDP. In particular, we require that the α-discounted stability coefficient

$$\beta_\psi \doteq \alpha \max_a \|P_a \psi\|_{\infty,\psi} \tag{2}$$

is strictly less than one. This can be seen to imply that $H : (\mathbb{R}^S, \|\cdot\|_{\infty,\psi}) \to (\mathbb{R}^{SA}, \|\cdot\|_{\infty,\psi})$ is a contraction, where for $J = (J_1^\top, \ldots, J_A^\top)^\top \in \mathbb{R}^{SA}$ we let $\|J\|_{\infty,\psi} = \max_a \|J_a\|_{\infty,\psi}$. That H is a contraction will play a crucial role in our results. Note that the condition $\beta_\psi < 1$ is closely related to the condition that for any policy u, $P_u \psi \le \psi$, which can be viewed as a stability condition on the MDP and which appeared in a slightly altered form in studying the stability of MDPs with infinite state spaces [e.g., 29]. Note that one can always choose $\psi = \mathbf{1}$, which gives $\beta_{\mathbf{1}} = \alpha < 1$. With this, we are ready to state our main result.

Theorem IV.1 (Error Bound for LRALP): Assume that $c \in \mathbb{R}_+^S$ is such that $1^\top c = 1$ and that $W \in \mathbb{R}_+^{SA \times m}$ is nonnegative valued. Let $\psi \in \mathbb{R}_+^S$ be in the column span of Φ and assume that the α-discounted stability coefficient of ψ is $\beta_\psi < 1$. Let $\varepsilon = \inf_{r \in \mathbb{R}^k} \|J^* - \Phi r\|_{\infty,\psi}$ be the error in approximating J^* using the basis functions in Φ. Then

$$\|J^* - J_{\mathrm{LRA}}\|_{1,c} \le \frac{2c^\top \psi}{1-\beta_\psi} \left(3\varepsilon + \|J^*_{\mathrm{ALP}} - J^*_{\mathrm{LRA}}\|_{\infty,\psi}\right).$$

Note that the result implicitly assumes that J_{LRA} exists, because if J_{LRA} does not exist then J^*_{LRA} is necessarily unbounded, making the last error term infinite. To ensure that ψ is in the span of Φ, after choosing ψ, one can add ψ as one of the basis functions. Alternatively, the bound can also be interpreted to hold for any ψ in the span of Φ with $\beta_\psi < 1$.

As noted earlier, de Farias and Van Roy [19] prove a similar error bound for J_{ALP}, the solution of the ALP. In particular, their Theorem 3 states that under identical assumptions as in our result, $\|J^* - J_{\mathrm{ALP}}\|_{1,c} \le \frac{2c^\top \psi \varepsilon}{1-\beta_\psi}$ for ε defined as above (the result we cited previously is a simplified form of this bound). The larger coefficient of ε is probably an artifact of our analysis. Note that when W does not reduce the constraints, our bound is only a constant factor larger than this previous result. The extra term $\|J^*_{\mathrm{ALP}} - J^*_{\mathrm{LRA}}\|_{\infty,\psi}$ can be seen as the price paid for relaxing the constraints.

From linear programming theory, it follows that primal boundedness is equivalent to dual feasibility. Since the dual of $\min\{c^\top x : Ax \ge b\}$ is $\max\{y^\top b : y \ge 0, c = A^\top y\}$, we get that a necessary and sufficient condition for J^*_{LRA} to be finite-valued is that for any $s \in \mathcal{S}$, $\phi(s)$ lies in the conic span, $\{U\lambda : \lambda \in \mathbb{R}_+^{SA}\}$, of (the columns) of $U = \Phi^\top E^\top W$. When W is such that its constituents $W_1, \ldots, W_A$ are all identical, the conic span of U is equal to the conic span of $\Phi^\top W_1$. It is particularly instructive to consider the case when the common matrix $W_a = (w_{1,a}, \ldots, w_{m,a})$ "selects" the m states, i.e., when $\{w_{1,a}, \ldots, w_{m,a}\} = \{e_s : s \in \mathcal{S}_0\}$ for some $\mathcal{S}_0 \subset \mathcal{S}$, $|\mathcal{S}_0| \le m$, where $e_s \in \{0,1\}^{\mathcal{S}}$ are the $s \in \mathcal{S}$ vectors in the standard Euclidean basis. In this case, the condition that $\phi(s)$ lies in the conic span of U is equivalent to that $\phi(s)$ lies in the conic span of $\phi(\mathcal{S}_0) \doteq \{\phi(s') : s' \in \mathcal{S}_0\}$. Thus, to ensure boundedness of J^*_{LRA}, the chosen states should be selected to "conicly cover" all the vectors in $\phi(\mathcal{S}) \subset \mathbb{R}^k$.[5]

The next theorem shows that magnitudes of the coefficients used in the conic cover control the size of $\|J^*_{\mathrm{ALP}} - J^*_{\mathrm{LRA}}\|_{\infty,\psi}$. For the theorem, we let $\Lambda \in \mathbb{R}_+^{\mathcal{S} \times \mathcal{S}_0}$ be the matrix of conic coefficients: For any $s \in \mathcal{S}$, $\phi(s) = \sum_{s' \in \mathcal{S}_0} \Lambda(s,s')\phi(s')$. After the theorem we give constructions for creating conic covers.

Theorem IV.2: Assume that $W_1 = \ldots = W_A$, $\{w_{1,a}, \ldots, w_{m,a}\} = \{e_s : s \in \mathcal{S}_0\}$ and that $\phi(\mathcal{S})$ lies in the conic span of $\phi(\mathcal{S}_0)$ with conic coefficients given by Λ. Let $\varepsilon = \inf_r \|J^* - \Phi r\|_{\infty,\psi}$. Then

$$\|J^*_{\mathrm{ALP}} - J^*_{\mathrm{LRA}}\|_{\infty,\psi} \le \|J^*_{\mathrm{ALP}} - J^*\|_{\infty,\psi} + (1 + \|\Lambda\psi\|_{\psi,\infty})\varepsilon\,.$$

Proof: Let r^* be such that $\|J^* - \Phi r^*\|_{\infty,\psi} = \varepsilon$ (this exists by continuity) and let $\delta = J^* - \Phi r^*$. Pick any $s \in \mathcal{S}$ and let $r_s = \mathrm{argmin}\{r^\top \phi(s) : W^\top E \Phi r \ge W^\top E J^*, r \in \mathbb{R}^k\}$ so that $J^*_{\mathrm{LRA}}(s) = r_s^\top \phi(s)$. Note that by assumption, for any $s' \in \mathcal{S}_0$, $J^*_{\mathrm{LRA}}(s') = r_s^\top \phi(s') \ge J^*(s')$. Now, notice that by definition, $J^*_{\mathrm{LRA}} \le J^*_{\mathrm{ALP}}$ (the LP defining J^*_{LRA} is the relaxation of the LP defining J^*_{ALP}). Hence, $0 \le J^*_{\mathrm{ALP}}(s) - J^*_{\mathrm{LRA}}(s) = J^*_{\mathrm{ALP}}(s) - J^*(s) + J^*(s) - J^*_{\mathrm{LRA}}(s)$ and $J^*_{\mathrm{LRA}}(s) = r_s^\top \phi(s) = r_s^\top \sum_{s' \in \mathcal{S}_0} \Lambda(s,s')\phi(s') = \sum_{s' \in \mathcal{S}_0} \Lambda(s,s') J^*_{\mathrm{LRA}}(s') \ge \sum_{s' \in \mathcal{S}_0} \Lambda(s,s') J^*(s')$. Combining this with the previous inequality we get $0 \le \frac{J^*_{\mathrm{ALP}}(s) - J^*_{\mathrm{LRA}}(s)}{\psi(s)} \le \frac{J^*_{\mathrm{ALP}}(s) - J^*(s)}{\psi(s)} + \frac{J^*(s) - \sum_{s' \in \mathcal{S}_0} \Lambda(s,s') J^*(s')}{\psi(s)}$. Plugging in $J^*(s) = \phi(s)^\top r^* + \delta(s)$, using again that $\phi(s) = \sum_{s' \in \mathcal{S}_0} \Lambda(s,s')\phi(s')$, and also using the triangle inequality after taking absolute values, we get

$$\begin{aligned}
&\frac{|J^*(s) - \sum_{s' \in \mathcal{S}_0} \Lambda(s,s') J^*(s')|}{\psi(s)} \\
&\quad \le \frac{|\delta(s)|}{\psi(s)} + \frac{\sum_{s' \in \mathcal{S}_0} \Lambda(s,s')|\delta(s')|}{\psi(s)} \\
&\quad \le \frac{|\delta(s)|}{\psi(s)} + \|\delta\|_{\infty,\psi} \frac{\sum_{s' \in \mathcal{S}_0} \Lambda(s,s')\psi(s')}{\psi(s)}\,.
\end{aligned}$$

Combining this with the previous display and noting that $\|\delta\|_{\infty,\psi} = \varepsilon$ finishes the proof. ■

Given $\phi : \mathcal{S} \to \mathbb{R}^k$, what is the minimum cardinality set $\mathcal{S}_0$ that conicly covers $\phi(\mathcal{S})$ and how to find such a set? Further, how to keep the magnitude of $\|\Lambda\psi\|_{\infty,\psi}$ small? To control this latter quantity, it seems essential to make sure $\mathcal{S}_0$ contains states with high ψ-values. However, if one is content with a bound that depends on $\|\psi\|_\infty$, one can bound $\|\Lambda\psi\|_{\infty,\psi}$ by $\|\psi\|_\infty \zeta$ where $\zeta = \max_s \sum_{s' \in \mathcal{S}_0} \Lambda(s,s')$, hence, the second term in the previous bound will be bounded by $(1 + \|\psi\|_\infty \zeta)\varepsilon$.

Let us now return to the problem of finding conic covers. We will proceed by considering some illustrative examples. As a start, consider the case when the basis functions are binary valued. In this case, it is sufficient and necessary to choose one state for each binary vector that appears in $\phi(\mathcal{S}) \subset \{0,1\}^k$. This gives that $m_0 \doteq |\mathcal{S}_0| \le 2^k$ representative states will be sufficient *regardless of the cardinality of* $\mathcal{S}$. Further, in this case $\zeta = 1$. For moderate to large k (e.g., $k \gg 20$), it will quickly become infeasible to keep 2^k constraints. In this case, we may need to restrict what features are considered to guarantee the conic cover condition. Letting $A_i = \{s \in \mathcal{S} : \phi_i(s) = 1\}$, if for a many pairs $i \ne j$, A_i, and A_j do not overlap then $N = |\phi(\mathcal{S})|$ can be much

[4] As opposed to de Farias and Van Roy [19] and others, our definition uses division and not multiplication with the weights. We choose this form for mathematical convenience: With this definition, nice duality results hold between weighted 1-norms and weighted maximum norms.

[5] The same implies that, under the same condition on W, boundedness of the LRALP holds if and only if $\sum_s c(s)\phi(s)$ is in the conic span of $\phi(\mathcal{S}_0)$. Note that this is easy to fulfill if the support of c has a small cardinality by adding all the states in the support of c to $\mathcal{S}_0$.

smaller than 2^k. For example, in the commonly used state aggregation procedures $A_i \cap A_j = \emptyset$ for any $i \neq j$, giving $N = k$. In the more interesting case of hierarchical aggregation (when the sets $\{A_i\}$ form a nested hierarchical partitioning of $\mathcal{S}$), we have $m_0 \leq D \cdot k$ where D is the depth of the hierarchy.

Another favourable example is the case of separable bases. In this case, the states are assumed to be factored and the basis functions depend only on a few factors. Let us consider a simple illustration. By abusing notation (redefining $\mathcal{S}$), let $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$, let there be $k = 2$ basis functions and assume that $\phi_i(s) = h_i(s)$ for some $h_i : \mathcal{S}_i \to \mathbb{R}$, $i = 1, 2$. Assume further that $0 \in h_i(\mathcal{S}_i)$ for both i and specifically let s_{i0} be such that $h_i(s_{i0}) = 0$. In this case, it is not hard to verify that if $\mathcal{S}_{i0}$ is such that $h_i(\mathcal{S}_i)$ is in the conic span of $h_i(\mathcal{S}_{i0})$, then $\phi(\mathcal{S})$ is also in the conic span of $\mathcal{S}_0 \doteq \mathcal{S}_1 \times \{s_{20}\} \cup \{s_{10}\} \times \mathcal{S}_2$. The point is that $|\mathcal{S}_0| \leq |\mathcal{S}_1| + |\mathcal{S}_2|$, which is a tolerable increase of growth. This example is not hard to generalize to more general, analysis of variance-like basis expansions. The moral is that as long as there is limited order of interaction (which is usually necessary for information theoretic reasons as well), the number of constraints may grow moderately with the number of factors (dimensionality) of the state space.

In some cases, finding a conic cover with a small cardinality is not possible. This can already happen in simple examples, such as when $\mathcal{S} = \{1, \ldots, S\}$ (as before) and $\phi(s) = (1, s, s^2)^\top$. In this case, the only choice is $\mathcal{S}_0 = \mathcal{S}$. In examples similar to this one, one possibility is to quantize the range of ϕ to trade a bit of the approximation quality away for reducing the cardinality of a conic cover.

The bound developed by de Farias and Van Roy [20] and our main result can be seen as largely complementary. Recall that de Farias and Van Roy [20] consider adding an extra constraint $r \in \mathcal{N}$, while they propose to select all A constraints from the ALP corresponding to m states chosen at random from some distribution μ. Then, with high probability, they show that, provided $r_{\text{ALP}} \in \mathcal{N}$, the extra price paid for relaxing the constraints of the ALP is $O(\rho\varepsilon_{\mathcal{N}} k/m)$, where $\rho = \max_s \frac{\mu^*(s)}{\mu(s)}$, $\mu^* = (1-\alpha)c^\top(I - \alpha P_{u^*})^{-1}$, u^* is an optimal policy, and $\varepsilon_{\mathcal{N}} = \sup_{r \in \mathcal{N}} \|J^* - \Phi r\|_{\infty,\psi}$.[6] The bound is nontrivial when $m \geq \rho\varepsilon_{\mathcal{N}} k$. In general, it may be hard to control ρ, or even $\varepsilon_{\mathcal{N}}$ while ensuring that $r_{\text{ALP}} \in \mathcal{N}$.

V. Proof of Theorem IV.1

In this section, we present the proof of the main result, Theorem IV.1. The proof uses contraction-arguments. We will introduce a novel contraction operator, $\hat{\Gamma} : \mathbb{R}^S \to \mathbb{R}^S$, that captures the distortion introduced by the extra constraint in ALP and the relaxation in LRALP, respectively. Then we relate the solution of LRALP to the fixed point of $\hat{\Gamma}$.

Note that for the proof, it suffices to consider the case when J^*_{LRA} is finite-valued because otherwise the bound is vacuous. Also, recall that it was assumed that ψ lies in the column space of Φ, while β_ψ, the α-discounted stability of ψ w.r.t. the MDP [cf., (2)] is strictly below one. We will let $r_0 \in \mathbb{R}^k$ be such that $\psi = \Phi r_0$. We also assumed that the matrix W is nonnegative valued, while c specifies a probability distribution over $\mathcal{S}$: $\sum_s c(s) = 1$ and $c \in \mathbb{R}^S_+$.

The operator $\hat{\Gamma}$ is defined as follows: For $J \in \mathbb{R}^S$, $s \in \mathcal{S}$

$$(\hat{\Gamma} J)(s) = \min\{r^\top \phi(s) \,:\, W^\top E \Phi r \geq W^\top E H J,\ r \in \mathbb{R}^k\}\,.$$

Note that $(\hat{\Gamma} J)(s)$ mimics the definition of ALP with $c = e_s$, except that the constraint $J = \Phi r$ is dropped.

[6]The paper presents the results for $\mu = \mu^*$ giving $\rho = 1$, but the analysis easily extends to the general case.

Let us now recall some basic results from the theory of contraction maps. First, let us recall the definition of contractions. Let $\|\cdot\|$ be a norm on $\mathbb{R}^S$ and $\rho > 0$. We say that the map $B : \mathbb{R}^S \to \mathbb{R}^S$ is $(\rho, \|\cdot\|)$-Lipschitz if for any $J, J' \in \mathbb{R}^S$, $\|BJ - BJ'\| \leq \rho\|J - J'\|$. We say that B is a $\|\cdot\|$-*contraction* with factor ρ if it is $(\rho, \|\cdot\|)$-Lipschitz and $\rho < 1$. It is particularly easy to check whether a map is a contraction map with respect to a weighted maximum norm if it is known to be *monotone*. Here, B is said to be monotone if for any $J \leq J'$, $J, J' \in \mathbb{R}^S$, $BJ \leq BJ'$ also holds, where $\leq$ is the componentwise partial order between vectors. We start with the following characterization of monotone contractions with respect to weighted maximum norms.

Lemma V.1: Let $B : \mathbb{R}^S \to \mathbb{R}^S$, $\psi : \mathcal{S} \to \mathbb{R}_{++}$, $\beta \in (0, 1)$. The following are equivalent:

1) B is a monotone contraction map with contraction factor β with respect to $\|\cdot\|_{\psi,\infty}$.
2) For any $J, J' \in \mathbb{R}^S$, $t \geq 0$, $J \leq J' + t\psi$ implies that $BJ \leq BJ' + \beta t\psi$.

Proof: The proof is a trivial modification of [16], proof of Lemma 3.1 and is thus left as an exercise. ■

Corollary V.2: If B is monotone and there exists some $\beta \in [0, 1)$ such that for any $J \in \mathbb{R}^S$ and $t > 0$,

$$B(J + t\psi) \leq BJ + \beta t\psi \tag{3}$$

then B is a $\|\cdot\|_{\infty,\psi}$ contraction with factor β.

Proof: Let $J, J' \in \mathbb{R}^S$, $t \geq 0$ and assume that $J \leq J' + t\psi$. By monotonicity $BJ \leq B(J' + t\psi)$, while by (3), $B(J' + t\psi) \leq BJ' + \beta t\psi$. Hence, $BJ \leq BJ' + \beta t\psi$. This shows that 2) of Lemma V.1 holds. Hence, by this lemma, B is a contraction with factor β with respect to $\|\cdot\|_{\infty,\psi}$. ■

Let us now return to the proof of our main result. Recall that the goal is to bound $\|J^* - J_{\text{LRA}}\|_{1,c}$ through relating this deviations from the fixed point of $\hat{\Gamma}$, which was promised to be a contraction. Let us thus now prove this. For this, it suffices to show that $\hat{\Gamma}$ satisfies the conditions of Corollary V.2. In fact, we will see this holds with $\beta = \beta_\psi$.

Proposition V.3: The operator $\hat{\Gamma}$ satisfies the conditions of Corollary V.2 with $\beta = \beta_\psi$, and is thus a $\|\cdot\|_{\infty,\psi}$-contraction with coefficient β_ψ.

Proof: First, note that (as it is well known) H is monotone (all the P_a matrices in the definition of H are nonnegative valued) and that it satisfies an inequality similar to (3): For any $t \geq 0$, $J \in \mathbb{R}^S$

$$H(J + t\psi) \leq HJ + \beta_\psi\, t\, E\psi\,. \tag{4}$$

This follows again because our assumption on ψ implies that for any $a \in \mathcal{A}$, $\alpha P_a \psi \leq \beta_\psi \psi$.

Let us now prove that $\hat{\Gamma}$ is monotone. Given $J \in \mathbb{R}^S$, let $\mathcal{F}'(J) \doteq \{\Phi r \,:\, W^\top E \Phi r \geq W^\top H J, r \in \mathbb{R}^k\}$. Choose any $s \in \mathcal{S}$. Since $J_1 \leq J_2$, W is nonnegative valued and H is monotone, we have $W^\top H J_1 \leq W^\top H J_2$. Hence, $\mathcal{F}_{J_2} \subset \mathcal{F}_{J_1}$ and thus $(\hat{\Gamma} J_1)(s) \leq (\hat{\Gamma} J_2)(s)$. Since s was arbitrary, monotonicity of $\hat{\Gamma}$ follows.

Let us now turn to proving that (3) holds with $\beta = \beta_\psi$. By definition, for $s \in \mathcal{S}$, $t \geq 0$, $J \in \mathbb{R}^S$, $(\hat{\Gamma}(J + t\psi))(s) = \min\{r^\top \phi(s) : W^\top E \Phi r \geq W^\top H(J + t\psi), r \in \mathbb{R}^k\}$. By (4), $H(J + t\psi) \leq HJ + t\beta_\psi E\psi$ and hence $W^\top H(J + t\psi) \leq W^\top(HJ + t\beta_\psi E\psi)$. Thus, $(\hat{\Gamma}(J + t\psi))(s) \leq \min\{r^\top \phi(s) \,:\, W^\top E\Phi r \geq W^\top(HJ + t\beta_\psi E\psi), r \in \mathbb{R}^k\}$.

To finish, we need the following elementary observation. ■

Claim V.4: Let $A \in \mathbb{R}^{u \times v}$, $b \in \mathbb{R}^u$, $d \in \mathbb{R}^v$ and $b_0 = Ax_0$ for some $x_0 \in \mathbb{R}^v$. Then

$$\min\{d^\top x : Ax \geq b + b_0, x \in \mathbb{R}^v\}$$
$$= \min\{d^\top y : Ay \geq b, y \in \mathbb{R}^v\} + d^\top x_0.$$

Proof of Claim V.4: Set $y = x - x_0$. ■

Now, using Claim V.4 with $A = W^\top E\Phi$, $b = W^\top HJ$, $d = \phi(s)$, $b_0 = t\beta_\psi W^\top E\psi$, and $x_0 = t\beta_\psi r_0$, thanks to $\Phi r_0 = \psi$ we have $Ax_0 = b_0$. Hence, the desired statement follows from the claim. ■

Let us now return to bounding $\|J^* - J_{\mathrm{LRA}}\|_{1,c}$. For $x \in \mathbb{R}$, let $(x)^-$ be the negative part of x: $(x)^- = \max(-x, 0)$. Then, $|x| = x + 2(x)^-$. For a vector $J \in \mathbb{R}^S$, we will write $(J)^-$ to denote the vector obtained by applying the negative part componentwise. We consider the decomposition

$$\|J_{\mathrm{LRA}} - J^*\|_{1,c} = c^\top (J_{\mathrm{LRA}} - J^*) + 2c^\top (J_{\mathrm{LRA}} - J^*)^- . \tag{5}$$

Let V_{LRA} be the fixed point of $\hat{\Gamma}$. We now claim the following.

Claim V.5: We have $J_{\mathrm{LRA}} \geq V_{\mathrm{LRA}}$, $c^\top J_{\mathrm{ALP}} \geq c^\top J_{\mathrm{LRA}}$.

Proof: The inequality $c^\top J_{\mathrm{ALP}} \geq c^\top J_{\mathrm{LRA}}$ follows immediately from the definitions of J_{ALP} and J_{LRA}.

To prove the first part, let $s \in \mathcal{S}$, $c = e_s$ and let r_s be a solution to LRALP in (1). For $s \in \mathcal{S}$, let $V_0(s) = \min_{s' \in \mathcal{S}} r_{s'}^\top \phi(s)$.

It suffices to show that $V_1 \doteq \hat{\Gamma} V_0 \leq V_0 \leq J_{\mathrm{LRA}}$. Indeed, if this holds then $V_{n+1} = \hat{\Gamma} V_n$, $n \geq 1$, satisfies $V_{n+1} \leq V_n$ and $V_n \to V_{\mathrm{LRA}}$ as $n \to \infty$ since $\hat{\Gamma}$ is a monotone contraction mapping.

Since $r_{s'}^\top \phi(s) \geq r_s^\top \phi(s)$ also holds for any $s, s' \in \mathcal{S}$, we have $V_0(s) = r_s^\top \phi(s)$. Also, since $J_{\mathrm{LRA}}(s) \geq r_s^\top \phi(s)$, it follows that $J_{\mathrm{LRA}} \geq V_0$. Now, fix some $s \in \mathcal{S}$ and let r'_{e_s, V_0} be the solution to the linear program defining $(\hat{\Gamma} V_0)(s)$. We need to show that $V_1(s) = (\hat{\Gamma} V_0)(s) = (r'_{e_s,V_0})^\top \phi(s) \leq V_0(s)$. By the definition of r'_{e_s,V_0}, we know that $(r'_{e_s,V_0})^\top \phi(s) \leq r^\top \phi(s)$ holds for any $r \in \mathbb{R}^k$ such that $W^\top E\Phi r \geq W^\top H V_0$. Thus, it suffices to show that r_s satisfies $W^\top E\Phi r_s \geq W^\top H V_0$. By definition, r_s satisfies $W^\top E\Phi r_s \geq W^\top H\Phi r_s$. Hence, by the monotone property of H and since W is nonnegative valued, it is sufficient if $\Phi r_s \geq V_0$. This, however, follows from the definition of V_0. ■

Thanks to the previous claim, $(J_{\mathrm{LRA}} - J^*)^- \leq (V_{\mathrm{LRA}} - J^*)^-$ and $c^\top J_{\mathrm{LRA}} \leq c^\top J_{\mathrm{ALP}}$. Hence, from (5) we get

$$\|J_{\mathrm{LRA}} - J^*\|_{1,c} \leq c^\top (J_{\mathrm{ALP}} - J^*) + 2c^\top (V_{\mathrm{LRA}} - J^*)^- .$$

By [19, Theorem 3], the first term is bounded by $2\frac{c^\top \psi}{1-\beta_\psi}\varepsilon$, where recall that $\varepsilon = \inf_r \|J^* - \Phi r\|_{\infty,\psi}$. Hence, it remains to bound the second term.

For this, note that for any $J \in \mathbb{R}^S$, $(J)^- \leq |J|$ and also that $\|J\|_{1,c} \leq c^\top \psi \|J\|_{\infty,\psi}$. Hence, we switch to bounding $\|J^* - V_{\mathrm{LRA}}\|_{\infty,\psi}$. A standard contraction argument gives

$$\begin{aligned}
\|J^* - V_{\mathrm{LRA}}\|_{\infty,\psi} &= \|J^* - \hat{\Gamma} J + \hat{\Gamma} J^* - \hat{\Gamma} V_{\mathrm{LRA}}\|_{\infty,\psi} \\
&\leq \|J^* - \hat{\Gamma} J^*\|_{\infty,\psi} + \|\hat{\Gamma} J^* - V_{\mathrm{LRA}}\|_{\infty,\psi} \\
&\leq \|J^* - \hat{\Gamma} J^*\|_{\infty,\psi} + \beta_\psi \|\hat{\Gamma} J^* - V_{\mathrm{LRA}}\|_{\infty,\psi} .
\end{aligned}$$

Reordering and using another triangle inequality we get

$$\|J^* - V_{\mathrm{LRA}}\|_{\infty,\psi} \leq \frac{\|J^* - J^*_{\mathrm{ALP}}\|_{\infty,\psi} + \|J^*_{\mathrm{ALP}} - \hat{\Gamma} J^*\|_{\infty,\psi}}{1 - \beta_\psi} .$$

We bound the term $\|J^* - J^*_{\mathrm{ALP}}\|_{\infty,\psi}$ in the following lemma.

Lemma V.6: We have $\|J^* - J^*_{\mathrm{ALP}}\|_{\infty,\psi} \leq 2\varepsilon$, where recall that $\varepsilon = \inf_{r \in \mathbb{R}^k} \|J^* - \Phi r\|_{\infty,\psi}$.

Proof: Let $r^* \doteq \operatorname{argmin}_{r \in \mathbb{R}^k} \|J^* - \Phi r\|_{\infty,\psi}$. First, notice that $J^*_{\mathrm{ALP}} \geq J^*$. Hence, $0 \leq J^*_{\mathrm{ALP}} - J^*$. Now let $r' = r^* + \varepsilon r_0$. Then, $\Phi r' = \Phi r^* + \varepsilon\psi \geq J^*$, where the equality follows by the definition of r_0 and the inequality follows by the definition of ε. Hence, r' is in the feasible set of the LP defining J^*_{ALP} and thus $J^*_{\mathrm{ALP}} \leq \Phi r'$. Thus, $0 \leq J^*_{\mathrm{ALP}} - J^* \leq \Phi r^* - J^* + \varepsilon\psi$. Dividing componentwise by ψ, taking absolute value and then taking maximum of both sides gives the result. ■

The proof of the main result is finished by noting that $\hat{\Gamma} J^* = J^*_{\mathrm{LRA}}$ and the chaining the inequalities we derived.

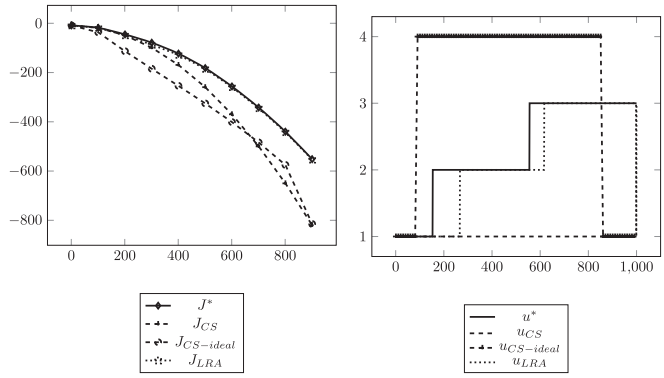

Fig. 1. Results for a single-queue with polynomial features. In both figures, the x-axis represents the state space: the length of the queue. For more details, see the text.

VI. Numerical Illustration

We consider a small scale queuing model similar to one in [19, Section 5.2]. The small scale helps to compare with the optimal policy. To stress-test the algorithms, we will use a correspondingly small LRALP. The queuing system has a single queue with random arrivals and departures. The state of the system is the queue length, an element of $\mathcal{S} = \{0, \ldots, S-1\}$. Action $a_t \in \mathcal{A} = \{1, \ldots, A\}$ sets service rates: Assuming that at time t the state is s_t, $s_{t+1} = \max(\min(s_t + \Delta_{t+1}, S-1), 0)$, where $\Delta_{t+1} \in \{-1, 0, +1\}$ with respective probabilities $q(a_t)$, $1 - q(a_t) - p$ and p. The service rates satisfy $0 < q(1) \leq \ldots \leq q(A) < 1$ with $q(A) > p$ so as to ensure "stabilizability" of the queue. The reward associated with action $a \in \mathcal{A}$ and state $s \in \mathcal{S}$ is given by $g_a(s) = -((s/N)^2 + q(a)^3)$. We ran our experiments[7] for $S = 1000$, $A = 4$ with $q(1) = 0.2$, $q(2) = 0.4$, $q(3) = 0.6$, $q(4) = 0.8$, $p = 0.2$, and $\alpha = 1 - \frac{1}{S}$. The ith basis function is $\phi_i(s) = s^i$, $i \in \{0, 1, \ldots, k-1\}$. Note that the conic span conditions will only be met with some lag, unless *all* constraints are selected. Hence, these features test the robustness of our theory. We chose $k = 4$, a low number, to counteract that the MDP is small scale. At any given state s, we formulate LRALPs to compute the approximate values $\hat{J}(s')$ at all the next states s' (except at the boundaries, each state has two next states). While formulating the LRALPs, we adopt different constraint selection strategies, which we call LRA, constraint sampling (CS), and CS-ideal, to be described below. Using the output of the LRALPs we compute *lookahead* policies as $u_\star(s) = \operatorname{argmin}_{a \in A} g_a(s) + \sum_{s' \in S} p_a(s, s') \hat{J}_\star(s')$, where $\star \in \{\mathrm{LRA}, \mathrm{CS}, \mathrm{CS\text{-}ideal}\}$. Each constraint selection strategy chooses all constraints corresponding to a select number of states $\mathcal{S}_0$. The strategies thus differ in how $\mathcal{S}_0$ is chosen. They also differ in the choice of the cost vector c. In LRA, we let $c = e_{s'}$ and set $\mathcal{S}_0 = \{1, 200, 400, 600, 800, 999, s'\}$. In CS, we let $c_{s'}(s) = \zeta(1-\alpha)(\alpha)^{|s'-s|}$ and choose $\zeta > 0$ so that $c_{s'}^\top \mathbf{1} = 1$. We then sample $m = 6$ states randomly from $c_{s'}$ to form $\mathcal{S}_0$. The idea is that $c_{s'}$ assigns higher weights to nearby states, while it provides a reasonable approximation to the "ideal" distribution that minimizes the upper bound of [20]. This distribution, $c_{s'} = e_{s'}^\top (1-\alpha)(I - \alpha P_{u^*})^{-1}$, is intractable (as u^* is not available), but was used in strategy CS-ideal to provide a strong baseline. The results are shown in the Fig. 1: The right-hand-side figure shows the policies computed, while the left-hand-side figure shows their value functions. Since the constraint sampling strategies produce randomized results, we repeated the computations ten times. The results in all cases were quite similar except

[7]The code can be found at https://tinyurl.com/y6w39jxq

for one "bad case" for CS. We show the plot for a "typical" run (not a bad one). As can be seen from the figure, choosing the constraints to (approximately) satisfy the constraint of the theoretical results reliably produces better results: In fact, the value functions J^* and J_{LRA} are mostly on the top of each other. While these results are only meant for illustration, we expect that in larger domains it becomes even more important to select the constraints in a systematic manner. The study of this is left for future work.

VII. Conclusion

In this paper, we introduced and analyzed the LRALP whose constraints were obtained as nonnegative linear combination of the original constraints of the ALP. The main novel contribution is a theoretical result that gives a geometrically interpretable bound on the performance loss due to relaxing the constraints. Possibilities for future work include extending the results to other forms of approximate linear programming in MDPs (e.g., [23]), exploring the idea of approximating dual variables and designing algorithms that use the newly derived results to actively compute what constraints to select.

References

[1] D. J. White, "A survey of applications of Markov decision processes," *J. Oper. Res. Soc.*, vol. 44, no. 11, pp. 1073–1096, 1993.

[2] J. Rust, "Numerical dynamic programming in economics," in *Handbook of Computational Economics*, vol. 1. North Holland, The Netherlands: Elsevier, 1996, pp. 619–729.

[3] E. A. Feinberg and A. Shwartz, *Handbook of Markov Decision Processes: Methods and Applications*. Norwell, MA, Kluwer , 2002.

[4] Q. Hu and W. Yue, *Markov Decision Processes With Their Applications*. Berlin, Germany: Springer, 2007.

[5] O. Sigaud and O. Buffet, Eds., *Markov Decision Processes in Artificial Intelligence*. Hoboken, NJ: USA: Wiley-ISTE, 2010.

[6] N. Bäuerle and U. Rieder, *Markov Decision Processes With Applications to Finance*. Berlin, Germany: Springer, 2011.

[7] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Programming*. New York, NY, USA: Wiley, 1994.

[8] F. L. Lewis and D. Liu, Eds., *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ: USA: Wiley, 2012.

[9] M. Abu Alsheikh, D. T. Hoang, D. Niyato, H.-P. Tan, and S. Lin, "Markov decision processes with applications in wireless sensor networks: A survey," *IEEE Comm. Surv. Tut.*, vol. 17, no. 3, pp. 1239–1267, Jul.–Sep. 2015.

[10] R. J. Boucherie and N. M. van Dijk, Eds., *Markov Decision Processes in Practice*, vol. 248. Berlin, Germany: Springer, 2017.

[11] J. Rust, "Using randomization to break the curse of dimensionality," *Econometrica*, vol. 65, pp. 487–516, 1996.

[12] C. Szepesvári, "Efficient approximate planning in continuous space Markovian decision problems," *AI Commun.*, vol. 13, no. 3, pp. 163–176, 2001.

[13] M. Kearns, Y. Mansour, and A. Y. Ng, "A sparse sampling algorithm for near-optimal planning in large Markov decision processes," *Mach. Learn.*, vol. 49, pp. 193–208, 2002.

[14] V. D. Blondel and J. N. Tsitsiklis, "A survey of computational complexity results in systems and control," *Automatica*, vol. 36, pp. 1249–1274, 2000.

[15] P. J. Schweitzer and A. Seidmann, "Generalized polynomial approximations in Markovian decision processes," *J. Math. Anal. Appl.*, vol. 110, pp. 568–582, 1985.

[16] L. Kallenberg, "Markov decision processes: Lecture notes," 2016. [Online]. Available: https://goo.gl/yhvrph

[17] D. Schuurmans and R. Patrascu, "Direct value-approximation for factored MDPs," in *Proc. Neural Inf. Process. Syst.*, 2001, pp. 1579–1586.

[18] C. Guestrin, D. Koller, R. Parr, and S. Venkataraman, "Efficient solution algorithms for factored MDPs," *J. Artif. Intell. Res.*, vol. 19, pp. 399–468, 2003.

[19] D. P. de Farias and B. Van Roy, "The linear programming approach to approximate dynamic programming," *Oper. Res.*, vol. 51, pp. 850–865, 2003.

[20] D. P. de Farias and B. Van Roy, "On constraint sampling in the linear programming approach to approximate dynamic programming," *Math. Oper. Res.*, vol. 29, pp. 462–478, 2004.

[21] B. Kveton and M. Hauskrecht, "Heuristic refinements of approximate linear programming for factored continuous-state Markov decision processes," in *Proc. Int. Conf. Automat. Planning Scheduling*, 2004, pp. 306–314.

[22] M. Petrik and S. Zilberstein, "Constraint relaxation in approximate linear programs," in *Proc. 26th Int. Conf. Mach. Learn.*, 2009, pp. 809–816.

[23] V. V. Desai, V. F. Farias, and C. C. Moallemi, "A smoothed approximate linear program," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 459–467.

[24] G. Taylor, M. Petrik, R. Parr, and S. Zilberstein, "Feature selection using regularization in approximate linear programs for Markov decision processes," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 871–878.

[25] J. Pazis and R. Parr, "Non-parametric approximate linear programming for MDPs," in *Proc. 25th AAAI Conf. Artif. Intell.*, 2011, pp. 459–464.

[26] N. Bhat, V. Farias, and C. C. Moallemi, "Non-parametric approximate dynamic programming via the kernel method," in *Proc. Neural Inf. Process. Syst. Conf.*, 2012, pp. 386–394.

[27] Y. Abbasi-Yadkori, P. Bartlett, and A. Malek, "Linear programming for large-scale Markov decision problems," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 496–504.

[28] C. Lakshminarayanan and S. Bhatnagar, "A generalized reduced linear program for Markov Decision Processes," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2722–2728.

[29] R.-R. Chen and S. P. Meyn, "Value iteration and optimization of multiclass queueing networks," *Queueing Systems*, vol. 32, no. 1-3, pp. 65–97, 1999.