A Generalized Reduced Linear Program for Markov Decision Processes

Chandrashekar Lakshminarayanan*, Shalabh Bhatnagar*, and Csaba Szepesvari†

***Abstract*—The approximate linear program (ALP) and its variants have been widely applied to Markov Decision Processes (MDPs) with large number of states. A serious limitation of the ALP is that it has intractable number of constraints, as a result of which constraint approximation is of interest. In this paper, we define a generalized reduced linear program (GRLP) that has a tractable number of constraints obtained as positive linear combinations of the original constraints of the ALP. The main contribution of this paper is a novel theoretical framework developed to obtain error bounds for any given GRLP. By providing a detailed error analysis for the GRLP, we justify linear approximation of the dual variables. Unlike prior probabilistic results on constraint sampling, our analysis is deterministic and is based on a novel contraction operator.**

Keywords: Approximate Dynamic Programming (ADP), Markov Decision Processes (MDPs), Approximate Linear Programming (ALP), Generalized Reduced Linear Program (GRLP), Constraint Sampling, Reinforcement Learning.

I. Introduction

Markov decision processes (MDPs) are a powerful mathematical framework to study optimal sequential decision making problems arising in science and engineering. The so-called dynamic programming methods find an optimal policy by computing the optimal *value-function* (J^*), a vector whose dimension is the number of states. MDPs with small number of states can be solved easily by conventional dynamic programming techniques, such as value-, or policy-iteration, or linear programming (LP) [1].

In this paper we consider the problem of using dynamic programming with MDPs with large state spaces. A practical way to tackle the issue of large number of states is to compute an approximate value function $\tilde{J}$ instead of J^*. Linear function approximation (LFA), i.e., letting $\tilde{J} = \Phi r^*$ where Φ is a so-called feature matrix and r^* is a weight vector to be computed, is the most widely used method of approximation. Here, dimensionality reduction is achieved by choosing Φ to have fewer columns in comparison to the number of states, holding the promise of being able to work with MDPs regardless of the number of states.

The *approximate linear program* (ALP) [3, 4, 5, 2] and its variants introduce linear function approximation in the linear programming formulation. A critical shortcoming of vanilla ALP is that the number of constraints are of the order of the size of the state space, making this vanilla version intractable. A way out is to choose a subset of constraints at random and drop the rest, thereby formulating a *reduced linear program* (RLP). The performance analysis of the RLP can be found in [4] and the RLP has also been shown to perform well in experiments [3, 4, 7]. An alternative approach to handle the issue of large number of constraints is to employ function approximation in the dual variables of the ALP [2, 6], an approach that was also found useful in experiments. However, to this date, there exist no theoretical guarantees bounding the loss in performance resulting from such an approximation.

In this paper, we generalize the RLP to define a generalized reduced linear program (GRLP) which has a tractable number of constraints that are obtained as positive linear combinations of the original constraints. The salient aspects of our contribution are listed below:

1) We develop novel analytical machinery to relate $\hat{J}$, the solution to the GRLP, and the optimal value function J^* by bounding the prediction error $||J^* - \hat{J}||$ (Theorem IV.15).
2) We also bound the performance loss due to using the policy $\hat{u}$ that is one-step greedy with respect to $\hat{J}$ (Theorem IV.16).
3) Our analysis is based on two novel $\max$-norm contraction operators and our results hold *deterministically*, as opposed to the results on RLP [5, 4], where the guarantees have a probabilistic nature.
4) Our results on the GRLP are the first to theoretically analyze the use of linear function approximation of Langrangian (dual) variables underlying the constraints.
5) A numerical example in controlled queues is provided to illustrate the theory.

A short and preliminary version of this paper without the theoretical analysis is available in [8].

II. Markov Decision Processes (MDPs)

We consider MDPs with a finite state space $S = \{1, 2, \ldots, n\}$ and finite action space $A = \{1, 2, \ldots, d\}$, where the sentiment is that n is large. For simplicity, we assume that all actions are feasible in all states. The probability transition kernel P collects the probabilities $p_a(s, s')$ of transitioning from state s to state s' under the action a for all possible $s, s' \in S$ and $a \in A$. We denote the reward (or gain) obtained for performing action $a \in A$ in state $s \in S$ by $g_a(s)$.

A stationary deterministic policy[1] (SDP), or simply a policy, is a map $u \colon S \to A$ that specifies for each state what action to select in that state. Given an SDP u, the expected total

*Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560012. E-mail: {chandrul, shalabh}@csa.iisc.ernet.in

†Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada T6G 2E8. E-mail: csaba.szepesvari@ualberta.ca

[1] For the scope of this paper, it suffices to restrict our attention to stationary deterministic policies.

discounted reward corresponding to starting at state s, while choosing actions as dictated by u for the states encountered, is

$$J_u(s) \triangleq \mathbf{E}[\sum_{n=0}^{\infty} \alpha^n g_{a_n}(s_n)|s_0 = s, a_n = u(s_n)\ \forall n \geq 0],$$

where $\alpha \in (0,1)$ is the so-called discount factor and $s_{n+1} \sim p_{a_n}(s_n, \cdot)$, $n \geq 0$. We call $J_u(s)$ the value of state s under SDP u, while J_u (as a map from S to the reals) is called the value function underlying policy u. The *optimal policy* u^* is one that in each state $s \in S$ achieves the best possible total expected discounted reward from that state. That is, $J^{u^*}(s) = J^*(s) \doteq \max_{u \in U} J_u(s)$ where U is the set of all SDPs and J^* is coined the *optimal value function*.[2],

Any optimal policy u^* and value function J^* obey the Bellman equation (BE): for all $s \in S$,

$$J^*(s) = \max_{a \in A} \big(g_a(s) + \alpha \sum_{s'} p_a(s,s')J^*(s')\big), \tag{1a}$$

$$u^*(s) = \operatorname*{argmax}_{a \in A} \big(g_a(s) + \alpha \sum_{s'} p_a(s,s')J^*(s')\big)\,. \tag{1b}$$

Once J^* is computed, u^* can be obtained via (1b) for each state relatively cheaply (e.g., having the ability to sample from the next-state distribution at a given state-action pair). Thus computing the value function is at the heart of most solution methods to MDP.

The value functions J^u or J^* are elements of R^S. In what follows it will be useful for us to treat these as n-dimensional vectors, i.e., elements of R^n, effectively identifying R^S with R^S in the natural way. Similarly we identify R^{nd} with $\mathrm{R}^{S \times A}$. The following definitions will be useful later:

Definition II.1. *Let $c, \rho, \chi : S \to \mathrm{R}_+$ be positive valued functions, where R_+ denotes the set of strictly positive reals. Then for $J \in \mathrm{R}^n$, $a \in A$ and $s \in S$, define*

(i) The Bellman operator $T\colon \mathrm{R}^n \to \mathrm{R}^n$ as $(TJ)(s) = \max_{a \in A}\big(g_a(s) + \alpha \sum_{s'} p_a(s,s')J(s')\big)$.

(ii) The Bellman operator (of action values) $H : \mathrm{R}^n \to \mathrm{R}^{nd}$ for state-action values as $HJ = [H_1 J, \cdots, H_d J]^\top \in \mathrm{R}^{nd}$, where $(H_a J)(s) = g_a(s) + \alpha \sum_{s'} p_a(s,s')J(s')$.

(iii) The weighted L_1-norms $\|\cdot\|_{1,c}$ and the weighted L_∞-norms $\|\cdot\|_{\infty,\rho}$ as $||J||_{1,c} = \sum_{s \in S} c(s)|J(s)|$, $||J||_{\infty,\rho} = \max_{s \in S} \frac{|J(s)|}{\rho(s)}$.

(iv) The discounted maximal inflation of χ due to $P = (p_a)_{a \in A}$ as $\beta_\chi = \max_{s \in S} \frac{\max_{a \in A}\left(\alpha \sum_{s'} p_a(s,s')\chi(s')\right)}{\chi(s)}$.

(v) The function $\chi : S \to \mathrm{R}_+$ above to be a Lyapunov *function for $P = (p_a)_{a \in A}$ if $\beta_\chi < 1$.*

(vi) E to be the $nd \times n$ matrix given by $E = [I, \ldots, I]^\top$, i.e., E is obtained by stacking d identical $n \times n$ identity matrices one over the other.

(vii) A policy $\tilde{u}$ is said to be greedy with respect to (w.r.t.) $\tilde{J} \in \mathrm{R}^n$ if for any $s \in S$,

$$\tilde{u}(s) \doteq \operatorname*{argmax}_{a \in A} \big(g_a(s) + \alpha \sum_{s'} p_a(s,s')\tilde{J}(s')\big).$$

We now state without proof the most important properties of the Bellman operator(s). The proofs are immediate from the definitions, but can also be found in [1]. First, we introduce some extra notation: For $J_1, J_2 \in \mathrm{R}^n$, we write $J_1 \leq J_2$ if $J_1(s) \leq J_2(s)$ holds for all $s \in S$. We use $\mathbf{1} \in \mathrm{R}^n$ to denote a vector with all entries 1. The maximum norm $\|\cdot\|_\infty$ is defined by $\|v\|_\infty = \max_{s \in S} |v(s)|$.

Lemma II.1. *T is a monotone map, i.e., given $J_1, J_2 \in \mathrm{R}^n$ such that $J_1 \leq J_2$, we have $TJ_1 \leq TJ_2$.*

Lemma II.2. *Given $J \in \mathrm{R}^n$ and $t \in \mathrm{R}$, we have*

$$T(J + t\mathbf{1}) = TJ + \alpha t \mathbf{1}. \tag{2}$$

Lemma II.3. *If $T : \mathrm{R}^n \to \mathrm{R}^n$ is any operator that is monotonous and satisfies (2) then T is a* max*-norm contraction operator with contraction factor $\alpha \in (0,1)$, i.e., given $J_1, J_2 \in \mathrm{R}^n$,*

$$||TJ_1 - TJ_2||_\infty \leq \alpha ||J_1 - J_2||_\infty. \tag{3}$$

Lemma II.4. *J^* is a unique fixed point of T, i.e., $J^* = TJ^*$.*

Corollary II.5. *If $J \in \mathrm{R}^n$ is such that $J \geq TJ$ then $J \geq TJ^2 \geq \ldots \geq J^*$.*

Though Lemmas II.1 to II.3 are stated for the Bellman operator T, the results also hold for H as well.

A. The Linear Programming Formulation

As it is well known, the optimal value function J^* can be obtained by solving the following linear program:

$$\min_{J \in \mathrm{R}^n} c^\top J \text{ s.t.} \tag{4}$$

$$J(s) \geq g_a(s) + \alpha \sum_{s'} p_a(s,s')J(s'),\ \forall (s,a) \in S \times A, \tag{5}$$

where c is *any* vector of R^n such that $c(s) > 0, \forall s \in S$. Without loss of generality, we may and will assume in what follows that $\sum_{s \in S} c(s) = 1$, i.e., c is a probability distribution over S.

When the MDP has a large number of states, it is difficult to solve for J^* using either the linear program (4) or other full state representation methods such as value iteration or policy iteration [1]. A practical solution is to resort to function approximation. Linear function approximation, wherein the solution is searched in the subspace spanned by a set of pre-selected basis functions, is an attractive choice.

III. Generalized Reduced Linear Program

The approximate linear program (ALP) is obtained by making use of LFA in the LP, i.e., by introducing the new variables $r \in \mathrm{R}^k$ and adding the extra constraint $J = \Phi r$ in (4) with $\Phi \in \mathrm{R}^{n \times k}$ [9]. By substitution, this leads to

$$\min_{r \in \mathrm{R}^k} c^\top \Phi r \ \text{ s.t. } \Phi r \geq T\Phi r, \tag{6}$$

where $J \geq TJ$ is a shorthand for the nd constraints in (4). Unless specified otherwise we use $\tilde{r}$ to denote an arbitrary solution to the ALP, and we let $\tilde{J} = \Phi\tilde{r}$ to denote the corresponding approximate value function and $\tilde{u}$ to denote

[2] In our case an optimal (SDP) u^* exists and is well defined [1].

the greedy policy w.r.t. $\tilde{J}$. The Generalized Reduced Linear Program is given as:

$$\min_{r\in\mathcal{N}} c^\top \Phi r \;\text{ s.t. }\; W^\top E\Phi r \geq W^\top H\Phi r, \tag{7}$$

where $W\in \mathrm{R}_+^{nd\times m}$ is a matrix with all positive entries and $\mathcal{N}$ is an additional (compact) constraint set to ensure the boundedness of the solution.[3] In what follows, we denote the solution to the GRLP by $\hat{r}$, the approximate value function by $\hat{J}=\Phi\hat{r}$ and use $\hat{u}$ to denote the greedy policy w.r.t. $\hat{J}$.

We assume that the following hold in the rest of the paper:

Assumption III.1. *(i) $c=(c(i),i=1,\dots,n)\in\mathrm{R}^n$ is a positive probability distribution, i.e., $c(i)>0\;\forall i$ and $\sum_{i=1}^n c(i)=1$.*

(ii) The first column of the feature matrix Φ (i.e., ϕ_1) is $\mathbf{1}\in\mathrm{R}^n$.

(iii) $\psi\colon S\to\mathrm{R}_+$ is a Lyapunov function for P and is present in the column span of the feature matrix Φ: For some $r_0\in\mathrm{R}^k$, $\Phi r_0=\psi$.

(iv) $\mathcal{N}\subset\mathrm{R}^k$ is compact and $\tilde{r}\in\mathcal{N}\subset\mathrm{R}^k$.

(v) $W\in\mathrm{R}_+^{nd\times m}$ is a full rank $nd\times m$ matrix (where $m\ll nd$), with all non-negative entries such that each of its column-sums equals one.

(vi) The set $\mathcal{N}'$ is such that $\mathcal{N}'=\mathcal{N}+tr_0$ for any $t\in\mathrm{R}$, where $r_0\in\mathrm{R}^k$ such that $\Phi r_0=\psi$.

We note in passing that if Assumption III.1-(iii) holds, it follows that for any $J\in\mathrm{R}^n$ and $t>0$,

$$T(J+t\psi)\leq TJ+\beta_\psi\, t\,\psi\,. \tag{8}$$

The GRLP introduces linear function approximation in both the primal and dual variables of the LP formulation. To understand this, we need to look at the Lagrangian of the ALP and GRLP in (9) and (10) respectively, i.e.,

$$\tilde{L}(r,\lambda)=c^\top\Phi r+\lambda^\top(T\Phi r-\Phi r), \tag{9}$$
$$\hat{L}(r,q)=c^\top\Phi r+q^\top W^\top(T\Phi r-\Phi r). \tag{10}$$

Thus, when $Wq=\lambda$, i.e., when W is a set of basis functions that allow a low dimensional linear representation of the dual variables λ, the two problems are the same. Hence, while the ALP employs LFA in its objective function (i.e., use of Φr), the GRLP employs linear approximation both in the objective function (Φr) as well as the constraints (use of W). To get a sense of how W should be chosen, recall that the optimal Lagrange multipliers are the discounted number of visits to the “state-action pairs” under an optimal policy u^*, i.e.,

$$\begin{aligned}\lambda^*(s,u^*(s))&=\big(c^\top(I-\alpha P_{u^*})^{-1}\big)(s)\\&=\big(c^\top(I+\alpha P_{u^*}+\alpha^2P_{u^*}^2+\dots)\big)(s),\\ \lambda^*(s,a)&=0,\qquad \text{for all } a\neq u^*(s),\end{aligned}$$

where P_{u^*} is the probability transition matrix under u^* ($P_{u^*}(s,s')=P_{u^*(s)}(s,s')$, $s,s'\in S$) [6]. Even though we might not have the optimal policy u^* in practice, the fact that λ^* is a probability distribution and that it is a linear combination of $\{P_{u^*},P_{u^*}^2,\dots\}$ hints at the kind of features that might be useful for the W matrix.

IV. Error Analysis

We now define two projection operators which are central to our error analysis:

Definition IV.1. *Given $J\in\mathrm{R}^n$ and the nonnegative valued vector $c\in\mathrm{R}_+^n$, define $r_{c,J}$ to be the solution to*

$$\min_{r\in\mathcal{N}'} c^\top\Phi r \;\text{ s.t. }\; \Phi r\geq TJ. \tag{11}$$

Then, for $J\in\mathrm{R}^n$, ΓJ, the least upper bound projection *of J is defined as*

$$(\Gamma J)(i)\doteq(\Phi r_{e_i,J})(i),\quad i=1,\dots,n\,. \tag{12}$$

Remark IV.1. *To understand the meaning of Γ (and $\hat{\Gamma}$) define*

$$\mathcal{F}_J\doteq\{\,\Phi r\,:\,\Phi r\geq TJ, r\in\mathcal{N}\,\}, \tag{13}$$

where $J\in\mathrm{R}^n$. Disregarding the constraint $r\in\mathcal{N}$, $\mathcal{F}_J$ contains vectors in the span of Φ that upper bound TJ. Further, since $(\Gamma J)(i)=\min\{V(i)\,:\,V\in\mathcal{F}_J\}$, it also follows that $V\geq\Gamma J$ holds for any $V\in\mathcal{F}_J$.

The definition of the second operator is as follows:

Definition IV.2. *Given $J\in\mathrm{R}^n$ and the nonnegative valued vector $c\in\mathrm{R}_+^n$, define $r'_{c,J}$ to be the solution to*

$$\min_{r\in\mathcal{N}'} c^\top\Phi r \;\text{ s.t. }\; W^\top E\Phi r\geq W^\top HJ. \tag{14}$$

Then, the approximate least upper bound (ALUB) projection operator *$\hat{\Gamma}\colon\mathrm{R}^n\to\mathrm{R}^n$ is defined as*

$$(\hat{\Gamma}J)(i)\doteq(\Phi r'_{e_i,J})(i),\;\; i=1,\dots,n\,,J\in\mathrm{R}^n\,. \tag{15}$$

The next lemma is elementary, but will prove to be useful:

Lemma IV.1. *Let $A\in\mathrm{R}^{u\times v}$, $b,c\in\mathrm{R}^u$ and $b_0=Ax_0$ for some $x_0\in\mathrm{R}^v$, $\mathcal{N}'\subset\mathrm{R}^v$ such that $\mathcal{N}'=x_0+\mathcal{N}'$. Then*

$$\begin{aligned}&\min\{c^\top Ax : Ax\geq b+b_0, x\in\mathcal{N}'\}\\&=\min\{c^\top Ay : Ay\geq b, y\in\mathcal{N}'\}+c^\top b_0.\end{aligned} \tag{16}$$

Proof. The claim follows by the change of variables $y:=x-x_0$. □

Lemma IV.2. *We have*

$$||J^*-\Gamma J^*||_{\infty,\psi}\leq 2||J^*-\Phi r^*||_{\infty,\psi}. \tag{17}$$

Proof. Define $\varepsilon=\|J^*-\Phi r^*\|_{\infty,\psi}$ so that $J^*-\Phi r^*\leq\varepsilon\psi$ and $\Phi r^*-J^*\leq\varepsilon\psi$. Then from Assumption III.1-(vi) it follows that $\Phi(r^*+\varepsilon r_0)=\Phi r^*+\varepsilon\psi\geq J^*=TJ^*$. From the definition of Γ in (12) and Remark IV.1, we know that $\Phi r^*+\varepsilon\psi\geq\Gamma J^*\geq J^*$. The result follows by noting that $2\varepsilon\psi\geq\Phi r^*+\varepsilon\psi-J^*\geq\Gamma J^*-J^*\geq 0$. □

Lemma IV.3. *For $J_1,J_2\in\mathrm{R}^n$ such that $J_1\leq J_2$, we have $\hat{\Gamma}J_1\leq\hat{\Gamma}J_2$.*

Proof. Given $J\in\mathrm{R}^n$, let $\mathcal{F}_J\doteq\{\,\Phi r\,:\,W^\top E\Phi r\geq W^\top HJ, r\in\mathcal{N}'\,\}$. Choose any $i\in\{1,\dots,n\}$. Since $J_1\leq J_2$, from Lemma II.1 and Assumption III.1-(v) it follows that $W^\top HJ_1\leq W^\top HJ_2$. Hence, $\mathcal{F}_{J_2}\subset\mathcal{F}_{J_1}$ and thus $(\hat{\Gamma}J_1)(i)\leq(\hat{\Gamma}J_2)(i)$. Since i was arbitrary, the result follows. □

[3] The appendix explains how $\mathcal{N}$ can be chosen.

Lemma IV.4. *Assume that $\hat{\Gamma} : \mathrm{R}^n \to \mathrm{R}^n$ is monotone and that there exists some $\beta \in [0, 1)$ such that for any $J \in \mathrm{R}^n$ and $t > 0$,*

$$\hat{\Gamma}(J + t\psi) \le \hat{\Gamma}J + \beta t\psi, \tag{18}$$

for any $J \in \mathrm{R}^n$ and $t \ge 0$. Then $\hat{\Gamma}$ is a $\| \cdot \|_{\infty,\psi}$ contraction with factor β.

Proof. First, we show that for any $t \ge 0$, $J \in \mathrm{R}^n$, $\hat{\Gamma}(J - t\psi) \ge \hat{\Gamma}J - \beta t\psi$ also holds. To see this define $J' = J - t\psi$. Then, $J = J' + t\psi$, hence $\hat{\Gamma}J \le \hat{\Gamma}J' + \beta t\psi$. Reordering this inequality gives the result. Let $\varepsilon = \|J_1 - J_2\|_{\infty,\psi}$, where $J_1, J_2 \in \mathrm{R}^n$ are arbitrary. Then $J_2 - \varepsilon\psi \le J_1 \le J_2 + \varepsilon\psi$. By the monotonicity of $\hat{\Gamma}$, $\hat{\Gamma}(J_2 - \varepsilon\psi) \le \hat{\Gamma}J_1 \le \hat{\Gamma}(J_2 + \varepsilon\psi)$. Using (18), we get $\hat{\Gamma}J_2 - \beta\varepsilon\psi \le \hat{\Gamma}J_1 \le \hat{\Gamma}J_2 + \beta\varepsilon\psi$, i.e., $-\beta\varepsilon\psi \le \hat{\Gamma}J_1 - \hat{\Gamma}J_2 \le \beta\varepsilon\psi$, from which the result follows. □

Corollary IV.5. *T is a $\| \cdot \|_{\infty,\psi}$-contraction with factor β_ψ.*

Lemma IV.6. *The operator $\hat{\Gamma}$ satisfies (18) with $\beta = \beta_\psi$.*

Proof. By definition, for $1 \le i \le n$, $(\hat{\Gamma}(J + t\psi))(i) = \min\{e_i^\top \Phi r \;:\; W^\top E\Phi r \ge W^\top H(J + t\psi), r \in \mathcal{N}'\}$. By (8), as $t > 0$, $H(J + t\psi) \le HJ + t\beta_\psi\psi$ and hence $W^\top H(J + t\psi) \le W^\top(HJ + t\beta_\psi\psi)$. Thus, $(\hat{\Gamma}(J + t\psi))(i) \le \min\{e_i^\top \Phi r \;:\; W^\top E\Phi r \ge W^\top HJ + t\beta_\psi\psi), r \in \mathcal{N}'\}$. Now, using Lemma IV.1 with $A = W^\top E\Phi$, $b = W^\top HJ$, $c = e_i$, $b_0 = t\beta_\psi\psi$ and $x_0 = t\beta_\psi r_0$, the statement follows since $Ax_0 = b_0$ (from Assumption III.1-(iii)) and $\mathcal{N}' = \mathcal{N}' + \alpha t r_0$ (from Assumption III.1-(vi)). □

Theorem IV.7. *The operator $\hat{\Gamma} : \mathrm{R}^n \to \mathrm{R}^n$ is a contraction operator in $\| \cdot \|_{\infty,\psi}$ with factor β_ψ.*

Proof. Follows from Lemmas IV.4 and IV.6. □

In what follows we denote by $\hat{V}$ the unique fixed point of $\hat{\Gamma}$, i.e., $\hat{V} = \hat{\Gamma}\hat{V}$. We now show that vector $\hat{J}$ dominates $\hat{V}$:

Lemma IV.8. *The vectors $\hat{V}, \hat{J}$ obey $\hat{J} \ge \hat{V}$.*

Proof. For $i \in \{1, \dots, n\}$, $c = e_i$ let r_i be a solution to the GRLP in (7) and define $V_0 \in \mathrm{R}^n$ by $V_0(i) = \min_{j=1,\dots,n}(\Phi r_j)(i)$, $1 \le i \le n$.

It suffices to show that $V_1 \doteq \hat{\Gamma}V_0 \le V_0 \le \hat{J}$ since then the desired result follows by defining $V_{n+1} = \hat{\Gamma}V_n$, $n \ge 1$, noting that by Lemma IV.3, $V_{n+1} \le V_n$ and by Corollary IV.5, $V_n \to \hat{V}$.

Since $(\Phi r_j)(i) \ge (\Phi r_i)(i)$ also holds for any $1 \le i, j \le n$ we have $V_0(i) = (\Phi r_i)(i)$. Also, since $\hat{J}(i) \ge (\Phi r_i)(i), 1 \le i \le n$ it follows that $\hat{J} \ge V_0$. Now, fix any i. We need to show that $V_1(i) = (\hat{\Gamma}V_0)(i) = (\Phi r'_{e_i,V_0})(i) \le V_0(i)$. By the definition of r'_{e_i,V_0} we know that $(\Phi r'_{e_i,V_0})(i) \le (\Phi r)(i)$ holds for any $r \in \mathcal{N}$ such that $W^\top E\Phi r \ge W^\top HV_0$. Now it suffices to show that r_i satisfies $W^\top E\Phi r_i \ge W^\top HV_0$. By definition, r_i satisfies $W^\top E\Phi r_i \ge W^\top H\Phi r_i$. Hence, by the monotone property of H and Assumption III.1-(v) it is sufficient if $\Phi r_i \ge V_0$. This however follows from the definition of V_0. □

Lemma IV.9. *The vectors $\hat{V}, \tilde{J}$ obey $\tilde{J} \ge \hat{V}$.*

Proof. Let $r_1, r_2, \dots, r_n$ be solutions to the ALP in (6) (with an additional constraint that the solution be restricted to $\mathcal{N}$) for $c = e_1, e_2, \dots, e_n$, respectively, and define $V_0 \in \mathrm{R}^n$ by $V_0(i) = \min_{j=1,\dots,n}(\Phi r_j)(i)$, $1 \le i \le n$. The rest of the proof follows in the same manner as the proof of Lemma IV.8. □

Lemma IV.10. *A vector $\hat{r} \in \mathrm{R}^k$ is a solution to GRLP (7) iff it solves the following program:*

$$\min_{r \in \mathcal{N}} ||\Phi r - \hat{V}||_{1,c} \;\; s.t. \;\; W^\top E\Phi r \ge W^\top H\Phi r. \tag{19}$$

Proof. We know from Lemma IV.9 that $\hat{J} = \Phi\hat{r} \ge \hat{V}$, and thus the solutions to (7) do not change if we add the constraint $\Phi r \ge \hat{V}$. Now, under this constraint, minimizing $c^\top \Phi r$ is the same as minimizing

$$||\Phi r - \hat{V}||_{1,c} = \sum_{i=1}^{n} c(i)|(\Phi r)(i) - \hat{V}(i)| = c^\top \Phi r - c^\top \hat{V}\,.$$

□

Lemma IV.11. *A vector $\tilde{r} \in \mathrm{R}^k$ is a solution to ALP (6) iff it solves the following program:*

$$\min_{r \in \mathrm{R}^k} ||\Phi r - \hat{V}||_{1,c} \;\; s.t. \;\; \Phi r \ge T\Phi r. \tag{20}$$

Proof. We know from Lemma IV.9 that $\tilde{J} = \Phi\tilde{r} \ge \hat{V}$. The rest of the argument follows in the same manner as the proof for Lemma IV.10. □

Lemma IV.12. *We have*

$$\begin{aligned} ||J^* - \hat{V}||_{\infty,\psi} \le \frac{1}{1 - \beta_\psi}\big(2||J^* - \Phi r^*||_{\infty,\psi} \\ + ||\Gamma J^* - \hat{\Gamma}J^*||_{\infty,\psi}\big). \end{aligned} \tag{21}$$

Proof. Recall that $\hat{\Gamma}\hat{V} = \hat{V}$. By the triangle inequality,

$$\begin{aligned} ||J^* - \hat{V}||_{\infty,\psi} &\le ||J^* - \hat{\Gamma}J^*||_{\infty,\psi} + ||\hat{\Gamma}J^* - \hat{\Gamma}\hat{V}||_{\infty,\psi} \\ &\le ||J^* - \hat{\Gamma}J^*||_{\infty,\psi} + \beta_\psi||J^* - \hat{V}||_{\infty,\psi}, \end{aligned}$$

and so by reordering and with another triangle inequality,

$$\begin{aligned} \|J^* - \hat{V}\|_{\infty,\psi} &\le \frac{\|J^* - \hat{\Gamma}J^*\|_{\infty,\psi}}{1 - \beta_\psi} \\ &\le \frac{\|J^* - \Gamma J^*\|_{\infty,\psi} + \|\Gamma J^* - \hat{\Gamma}J^*\|_{\infty,\psi}}{1 - \beta_\psi}\,. \end{aligned} \tag{22}$$

The proof follows by applying Lemma IV.2 on (22). □

We now recall Lemma 5 from Section 4.2 of [3]. For this result, recall that $r_0 \in \mathrm{R}^k$ is the vector such that $\psi = \Phi r_0$.

Lemma IV.13. *For $r \in \mathrm{R}^k$ arbitrary, let r' be*

$$r' = r + ||J^* - \Phi r||_{\infty,\psi}\left(\frac{1 + \beta_\psi}{1 - \beta_\psi}\right) r_0. \tag{23}$$

Then r' is feasible for the ALP in (6).

Recall that $\hat{V}$ is the fixed point of $\hat{\Gamma}$ and $\hat{J} = \Phi\hat{r}$ is the solution to the GRLP (7).

Theorem IV.14. *We have*

$$||\hat{J} - \hat{V}||_{1,c} \le \frac{c^\top\psi}{1 - \beta_\psi}(4||J^* - \Phi r^*||_{\infty,\psi} + ||\Gamma J^* - \hat{\Gamma}J^*||_{\infty,\psi}).$$

Proof. Let $\gamma = ||J^* - \Phi r^*||_{\infty,\psi}$. Then, by choosing r' as in Lemma IV.13 we have

$$\begin{aligned} ||\Phi r' - J^*||_{\infty,\psi} &\leq ||\Phi r^* - J^*||_{\infty,\psi} + ||\Phi r' - \Phi r^*||_{\infty,\psi} \\ &= \gamma + \frac{1+\beta_\psi}{1-\beta_\psi}\gamma = \frac{2}{1-\beta_\psi}\gamma. \end{aligned} \tag{24}$$

Now, r' is feasible for the ALP in (7) by Lemma IV.13. Then from Lemma IV.11 it follows that

$$\begin{aligned} ||\hat{J} - \hat{V}||_{1,c} &\leq ||\Phi\tilde{r} - \hat{V}||_{1,c} \leq ||\Phi r' - \hat{V}||_{1,c} \\ &= \sum_{s\in S} c(s)\psi(s)\frac{|\Phi r'(s) - \hat{V}(s)|}{\psi(s)} \\ &\leq c^\top\psi||\Phi r' - \hat{V}||_{\infty,\psi} \\ &\leq c^\top\psi(||\Phi r' - J^*||_{\infty,\psi} + ||J^* - \hat{V}||_{\infty,\psi}). \end{aligned}$$

The result follows from Lemma IV.12 and (24). □

Theorem IV.15 (Prediction error bound in $||\cdot||_{\infty,\psi}$). *It holds that*

$$\begin{aligned} ||J^* - \hat{J}||_{1,c} \leq \frac{c^\top\psi}{1-\beta_\psi}&(6||J^* - \Phi r^*||_{\infty,\psi} \\ &+ 2||\Gamma J^* - \hat{\Gamma}J^*||_{\infty,\psi}). \end{aligned} \tag{25}$$

Proof. We have

$$\begin{aligned} ||J^* - \hat{J}||_{1,c} &\leq ||J^* - \hat{V}||_{1,c} + ||\hat{V} - \hat{J}||_{1,c} \\ &\leq c^\top\psi||J^* - \hat{V}||_{\infty,\psi} + ||\hat{V} - \hat{J}||_{1,c}. \end{aligned}$$

The result now follows from Lemma IV.12 and Theorem IV.14. □

We now bound the performance of the greedy policy $\hat{u}$.

Theorem IV.16 (Control Error Bound in $||\cdot||_{\infty,\psi}$). *Let $\hat{u}$ be the greedy policy with respect to the solution $\hat{J}$ of the GRLP and $J_{\hat{u}}$ be its value function. Then,*

$$\begin{aligned} ||J^* - J_{\hat{u}}||_{1,c} \leq 2\left(\frac{c^\top\psi}{(1-\beta_\psi)^2}\right)&\big(2||J^* - \Phi r^*||_{\infty,\psi} \\ &+ ||\Gamma J^* - \hat{\Gamma}J^*||_{\infty,\psi} + ||\hat{J} - \hat{\Gamma}\hat{J}||_{\infty,\psi}\big). \end{aligned} \tag{26}$$

Proof. By the triangle inequality,

$$||J^* - J_{\hat{u}}||_{1,c} \leq ||J^* - \hat{J}||_{1,c} + ||J_{\hat{u}} - \hat{J}||_{1,c}\,.$$

Let us now bound the second term on the right-hand side. Since $\hat{u}$ is greedy w.r.t. $\hat{J}$, it holds that $T_{\hat{u}}\hat{J} = T\hat{J}$. Also, $T_{\hat{u}}J_{\hat{u}} = J_{\hat{u}}$. Hence, $J_{\hat{u}} - \hat{J} = T_{\hat{u}}J_{\hat{u}} - T_{\hat{u}}\hat{J} + T\hat{J} - \hat{J} = \alpha P_{\hat{u}}(J_{\hat{u}} - \hat{J}) + T\hat{J} - \hat{J}$. Hence,

$$\begin{aligned} ||J_{\hat{u}} - \hat{J}||_{1,c} &= ||(I - \alpha P_{\hat{u}})^{-1}(T\hat{J} - \hat{J})||_{1,c} \\ &\leq c^\top(I - \alpha P_{\hat{u}})^{-1}|T\hat{J} - \hat{J}| \\ &\leq c^\top(I - \alpha P_{\hat{u}})^{-1}\,\psi\,||T\hat{J} - \hat{J}||_{\infty,\psi} \\ &\leq \frac{c^\top\psi}{1-\beta_\psi}||T\hat{J} - \hat{J}||_{\infty,\psi} \\ &\leq \frac{c^\top\psi}{1-\beta_\psi}(||T\hat{J} - TJ^*||_{\infty,\psi} + ||J^* - \hat{J}||_{\infty,\psi}) \\ &\leq \frac{c^\top\psi}{1-\beta_\psi}(1+\beta_\psi)||J^* - \hat{J}||_{\infty,\psi}, \end{aligned} \tag{27}$$

where in the second inequality, we use Jensen's inequality and $|T\hat{J} - \hat{J}|$ stands for the vector whose ith component is $|(T\hat{J})(i) - \hat{J}(i)|$. Further, the last inequality follows since T is a $||\cdot||_{\infty,\psi}$ contraction with factor β_ψ as noted earlier. Hence,

$$\begin{aligned} &||J^* - J_{\hat{u}}||_{1,c} \\ &\leq c^\top\psi||J^* - \hat{J}||_{\infty,\psi} + c^\top\psi\frac{1+\beta_\psi}{1-\beta_\psi}||J^* - \hat{J}||_{\infty,\psi} \\ &= \frac{2c^\top\psi}{1-\beta_\psi}||J^* - \hat{J}||_{\infty,\psi}. \end{aligned} \tag{28}$$

Now in a manner similar to Theorem IV.15 we have

$$||J^* - \hat{J}||_{\infty,\psi} \leq ||J^* - \hat{V}||_{\infty,\psi} + ||\hat{V} - \hat{J}||_{\infty,\psi}$$

The result now follows by substituting the bound on $||J^* - \hat{V}||_{\infty,\psi}$ from Lemma IV.12 and the fact that $||\hat{V} - \hat{J}||_{\infty,\psi} \leq \frac{1}{1-\beta_\psi}||\hat{J} - \hat{\Gamma}\hat{J}||$. □

V. Discussion

The error bounds in the main results (Theorems IV.15 and IV.16) contain two factors, namely

1) $\min_{r\in\mathbb{R}^k}||J^* - \Phi r||_{\infty,\psi}$, and
2) $||\Gamma J^* - \hat{\Gamma}J^*||_{\infty,\psi}$.

The first factor is related to the best possible approximation that can be achieved with the chosen feature matrix Φ. This term is inherent to the ALP formulation and it appears in the bounds provided by [3].

The second factor is related to constraint approximation and is completely defined in terms of Φ, W and T, and does not require knowledge of stationary distribution of the optimal policy. It makes intuitive sense since given that Φ approximates J^*, it is enough for W to depend on Φ and T without any additional requirements.

An interesting feature is that unlike prior work on constraint sampling based on concentration inequalities (e.g., [4]), our analysis is based on contraction operators and is completely deterministic. In particular, the error term $||\Gamma J^* - \hat{\Gamma}J^*||_{\infty,\psi}$ gives new insights into constraint selection:

Theorem V.1. *Let $s \in S$ be a state whose constraint is selected by W (i.e., for some i and all $(s',a) \in S \times A$, $W_{s'a,i} = \delta_{s=s'}$. Then*

$$|\Gamma J^*(s) - \hat{\Gamma}J^*(s)| < |\Gamma J^*(s) - J^*(s)|. \tag{29}$$

Proof. Let r_{e_s,J^*} and r'_{e_s,J^*} be solutions to the linear programs in (11) and (14) respectively for $c = e_s$ and $J = J^*$. It is easy to note that r_{e_s,J^*} is feasible for the linear program in (14) for $c = e_s$ and J^*, and hence it follows that $(\Phi r_{e_s,J^*})(s) \geq (\Phi r'_{e_s,J^*})(s)$. However, since the constraints with respect to state s have been chosen we know that $(\Phi r'_{e_s,J^*})(s) \geq J^*(s)$. The proof follows from noting that $(\Gamma J^*)(s) = (\Phi r_{e_s,J^*})(s)$ and $\hat{\Gamma}J^*(s) = (\Phi r_{e_s,J^*})(s)$. □

The expression in (29) in Theorem V.1 says that the additional error $|\Gamma J^*(s) - \hat{\Gamma}J^*(s)|$ due to constraint approximation is less than the original projection error $|\Gamma J^*(s) - J^*(s)|$ due to function approximation. This means that for the RLP to perform well it is enough to retain those states for which the linear function approximation via Φ is known to perform well. The modified L_∞ norm in (25) comes to our rescue to control the error due to those states that are not chosen.

VI. Numerical Illustration

In this section, we show via an example in the domain of controlled queues that the error term $||\Gamma J^* - \hat{\Gamma} J^*||_\infty$ indeed correlates with the error induced by the constraint approximation. The queuing model used here is similar to the one in Section 5.2 of [3]. We consider a single queue with arrivals and departures. The state of the system is the queue length with the state space given by $S = \{0, \ldots, n-1\}$, where $n-1$ is the buffer size of the queue. The action set $A = \{1, \ldots, d\}$ is related to the service rates. We let s_t denote the state at time t. The state at time $t+1$ when action $a_t \in A$ is chosen is given by $s_{t+1} = s_t + 1$ with probability p, $s_{t+1} = s_t - 1$ with probability $q(a_t)$ and $s_{t+1} = s_t$, with probability $(1-p-q(a_t))$. For states $s_t = 0$ and $s_t = n-1$, the system dynamics is given by $s_{t+1} = s_t + 1$ with probability p when $s_t = 0$ and $s_{t+1} = s_t - 1$ with probability $q(a_t)$ when $s_t = n-1$. The service rates satisfy $0 < q(1) \leq \ldots \leq q(d) < 1$ with $q(d) > p$ so as to ensure 'stabilizability' of the queue. The reward associated with the action $a \in A$ in state $s \in S$ is given by $g_a(s) = -(s + 60q(a)^3)$. We made use of polynomial features in Φ (i.e., $1, s, \ldots, s^{k-1}$) since they are known to work well for this domain [3]. For our experiments, we chose two contenders for the W-matrix and compared them with random positive matrices W_r. Our choices of the W matrix were: (**i**) W_c- matrix that corresponds to sampling according to c. This is justified by the insights obtained from Theorem V.1 on the error term $||\Gamma J^* - \hat{\Gamma} J^*||_\infty$, i.e., the idea of selecting the important states. (**ii**) W_a state-aggregation matrix, a heuristic derived using the fact that λ^* is a linear combination of $\{P_{u^*}, P^2_{u^*}, \ldots\}$. Our choice of the W_a matrix to correspond to aggregation of nearby states is motivated by the observation that P^n captures n^{th} hop connectivity/neighborhood information. The aggregation matrix W_a is defined as below: for all $i = 1, \ldots, m$,

$$\begin{aligned} W_a(i,j) = 1, &\text{ for all } j \text{ s.t.} \\ & j = (i-1) \times \frac{n}{m} + k + (l-1) \times n, \\ & \quad k = 1, \ldots, \frac{n}{m}, l = 1, \ldots, d, \\ & = 0, \text{ otherwise.} \end{aligned} \tag{30}$$

We ran our experiments on a moderately large queuing system denoted by Q_L with $n = 10^4$ and $d = 4$ with $q(1) = 0.2$, $q(2) = 0.4$, $q(3) = 0.6$, $q(4) = 0.8$, $p = 0.4$ and $\alpha = 0.98$. We chose $k = 4$ (i.e., we used $1, s, s^2$ and s^3 as basis vectors) and we chose W_a (30), W_c, W_i and W_r with $m = 50$, where W_i is the ideal (and intractable) sampler of [4]. We set $c(s) = (1-\zeta)\zeta^s$, where $s = 1, \ldots, 9999$, with $\zeta = 0.9$ and $\zeta = 0.999$ respectively. The results in Table I show that the performance exhibited by W_a and W_c is better by several orders of magnitude over 'random' in the case of the large system Q_L and is closer to the ideal sampler W_i. Note that computing $||\Gamma J^* - \hat{\Gamma} J^*||_\infty$ was hard in the case of large $n = 10^4$ and since $||\Gamma J^* - \hat{\Gamma} J^*||_\infty$ is completely dependent on the structure of Φ, T and W we instead computed the same for $n = 10$ and used it as a surrogate. Accordingly, we chose a smaller system Q_S with $n = 10$, $d = 2$, $k = 2$, $m = 5$, $q(1) = 0.2$, $q(2) = 0.4$, $p = 0.2$ and $\alpha = 0.98$. Note that a better performance of W_a and W_c is in agreement with a lower value of $||\Gamma J^* - \hat{\Gamma} J^*||_\infty$.

Error Terms	W_i	W_c	W_a	W_r
$\|\|J^* - \hat{J}\|\|_{1,c}$ for $\zeta = 0.9$	32	32	220	5.04×10^4
$\|\|J^* - \hat{J}\|\|_{1,c}$ for $\zeta = 0.999$	110	180.5608	82	1.25×10^7
$\|\|\Gamma J^* - \hat{\Gamma} J^*\|\|_\infty$	0	39	24	214

TABLE I

Shows values of Error Terms for Q_L.

VII. Conclusion

In this paper, we introduced and analyzed the generalized reduced linear program (GRLP) whose constraints were obtained as positive linear combination of the original constraints of the ALP. By providing performance analysis for the GRLP, this paper justified linear function approximation of the dual variables of the ALP.

Appendix

On choice of $\mathcal{N}$:

Let $g_{\min} = \min\limits_{a,s} g_a(s)$, $g_{\max} = \max\limits_{a,s} g_a(s)$ and define $r^* \doteq \arg\min\limits_{r \in \mathrm{R}^k} ||J^* - \Phi r||_{\infty,\psi}$. Now $\frac{g_{\min}}{1-\alpha} \leq J^* \leq \frac{g_{\max}}{1-\alpha}$ and since $\mathbf{1}$ is in the basis, we have $||J^* - \Phi r^*||_{\infty,\psi} \leq \frac{g_{\max} - g_{\min}}{1-\alpha}$ (making use of the fact that for any sensible choice of Lyapunov function $||J^* - \Phi r^*||_{\infty,\psi} \leq ||J^* - \Phi r^*||_\infty$). Further, in our arguments, we have made the implicit assumption that $r' = r^* + ||J^* - \Phi r^*||_{\infty,\psi} \left(\frac{1+\beta_\psi}{1-\beta_\psi}\right) r_0 \in \mathcal{N}$. This implicit assumption will continue to hold when $\mathcal{N} \doteq \{\mathbf{r} | a \leq \Phi r \leq b\}$, where $a = \frac{g_{\min}}{1-\alpha}$ and $b = \frac{g_{\max}}{1-\alpha} + \frac{1+\beta_\psi}{1-\beta_\psi} \frac{g_{\max} - g_{\min}}{1-\alpha}$.

References

[1] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmont, MA, 4th edition, 2013. 1, 2

[2] V. S. Borkar, J. Pinto, and T. Prabhu. A new learning algorithm for optimal stopping. *Discrete Event Dynamic Systems*, 19(1):91–113, 2009. 1

[3] D. P. de Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003. 1, 4, 5, 6

[4] D. P. de Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, 29(3):462–478, 2004. 1, 5, 6

[5] V. V. Desai, V. F. Farias, and C. C. Moallemi. A smoothed approximate linear program. In *NIPS*, pages 459–467, 2009. 1

[6] D. A. Dolgov and E. H. Durfee. Symmetric approximate linear programming for factored MDPs with application to constrained problems. *Annals of Mathematics and Artificial Intelligence*, 47(3-4):273–293, August 2006. 1, 3

[7] V. F. Farias and B. Van Roy. Tetris: A study of randomized constraint sampling. In *Probabilistic and Randomized Methods for Design Under Uncertainty*, pages 189–201. Springer, 2006. 1

[8] C. Lakshminarayanan and S. Bhatnagar. A generalized reduced linear program for Markov Decision Processes. *Proceedings of Association for the Advancement of Artificial Intelligence (AAAI) Austin, Texas*, 2015. 1

[9] P. J. Schweitzer and A. Seidmann. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110:568–582, 1985. 2