A Generalized Reduced Linear Program for Markov Decision Processes

Chandrashekar Lakshminarayanan and Shalabh Bhatnagar
Department of Computer Science and Automation
Indian Institute of Science
Bangalore-560012
{chandrul,shalabh}@csa.iisc.ernet.in
and
Csaba Szepesvári
Department of Computing Science
University of Alberta
Edmonton, Alberta
Canada T6G 2E8
csaba.szepesvari@ualberta.ca

Abstract

The term Approximate Dynamic Programming (ADP) refers to a gamut of approximate solution methods for MDPs with large number of states. Though various ADP algorithms are known till date, very few of them successfully address both the prediction and the control problems. Approximate Linear Programming (ALP) is an ADP method that offers sound theoretical guarantees and solves both the prediction and the control problems. Nevertheless, ALP has a serious limitation in that it has large number of constraints, and in practice, a reduced linear program (RLP) is solved instead. Though the RLP has been shown to perform well empirically, error bounds are available only for a specific RLP obtained under idealized assumptions.

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 of the ALP. The main contribution of this paper is the novel theoretical framework developed

to obtain error bounds for any given GRLP. Central to our framework are two max-norm contraction operators. Our result also theoretically justifies linear approximation of constraints. We discuss the implication of our results in the contexts of ADP and reinforcement learning.

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

I. Introduction

Optimal sequential decision making problems occurring in science, engineering and economics can be cast in the framework of Markov Decision Processes (MDPs), where the problem is to find a policy u, mapping states to actions, so as to maximize long term expected cumulated discounted reward. A given policy u is associated with a value function J_u (a mapping from the state space to the set of reals[1]), which gives the value of each state under a given policy u. The optimal value function J^* collects the highest values achievable in each state. A policy u^* is optimal if it achieves the optimal value in each state, i.e., if $J^* = J_{u^*}$. Conventional solution methods for Markov Decision Processes (MDPs) such as value iteration, policy iteration and linear programming formulation compute the optimal value function and the optimal policy. However, computing the exact quantities for each and every state is hard when there are a large number of states.

A practical way to address the issue of large number of states is function approximation, wherein, the idea is to choose a function belonging to a parameterized family as an approximation to the exact value function. Linear function approximation is the most common, wherein, the value function is approximated as $J^* \approx \tilde{J} = \Phi r^*$, where Φ is a feature matrix whose columns are the basis functions and r^* is a weight vector to be learned. A representational advantage (in terms of the number of unknowns) is achieved by choosing fewer number of basis functions compared to the number of states. Once approximate value function $\tilde{J}$ is known, a sub-optimal policy $\tilde{u}$ which is greedy with respect to $\tilde{J}$ can be computed.

The *approximate linear programming* (ALP) [4, 5, 6, 3, 9, 13, 11] host of methods introduce linear function approximation in the linear programming formulation. A critical shortcoming of ALP is that the number of constraints are of the order of the state space, making, in the

[1]Also known as reward/cost-to-go function.

lack of extra structure, the vanilla version of ALP intractable. One proposal in the literature to overcome this hurdle is to employ a procedure known as constraint sampling, wherein a subset of the original constraints of the ALP are sampled to formulate a *reduced linear program* (RLP). The perfromance analysis of the RLP can be found in [5]. The RLP has been shown to perform well in experiments [4, 5, 8] in various domains such as Tetris and in network of queues. An alternative approach to handle the constraints is employ function approximation in the dual variables of the ALP [3]. However, [3] does not provide theoretical guarntees for the loss in performance due to such an approximation.

The success of the RLP [5, 8]and the idea of approximating the dual variables [3], naturally leads us to the question understanding linear function approximation of the constraints. To this end, 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 of the ALP. 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^* (Theorem V.13).
2) We also bound (Theorem V.14) the loss $||J^* - J_{\hat{u}}||$ of the one-step greedy policy $\hat{u}$ based on $\hat{J}$.
3) Our analysis is based on a novel $\max$-norm contraction operator and our results hold with probability one. This is another significant difference in comparison to the results on constraint sampling in [6, 5] that make use of concentration bounds and hold only with *high* probability.
4) The structure of the error terms also reveals that it is not always necessary to sample using the stationary distribution of the optimal policy.
5) Our results on the GRLP are the first to theoretically justify linear function approximation of the constraints.

II. Markov Decision Processes (MDPs)

In this section, briefly discuss the basics of Markov Decision Processes (MDPs) (the reader is referred to [2, 12] for a detailed treatment).

We consider MDPs with large but finite number of states, i.e., $S = \{1, 2, \ldots, n\}$ for some large n, and the action set is given by $A = \{1, 2, \ldots, d\}$. For simplicity, we assume that all actions are feasible in all states. The probability transition kernel P specifies the probability $p_a(s, s')$

of transitioning from state s to state s' under the action a. We denote the reward obtained for performing action $a \in A$ in state $s \in S$ by $g_a(s)$.

A stationary deterministic policy[2](SDP) or simply a policy is a map $u\colon S \to A$ which specifies the action selection mechanism. Given an SDP u, the infinite horizon discounted reward corresponding to state s under u is denoted by $J_u(s)$ and is defined by

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

where $\alpha \in (0,1)$ is a given discount factor. Here $J_u(s)$ is known as the value of the state s under the SDP u, and the vector quantity $J_u \stackrel{\Delta}{=} (J_u(s), \forall s \in S) \in R^n$ is called the value-function corresponding to the SDP u. The *optimal policy* u^* is obtained as $u^*(s) \stackrel{\Delta}{=} \arg\max_{u \in U} J_u(s)$[3], where U is the class of all SDPs. The *optimal value-function* J^* is the one obtained under the optimal policy, i.e., $J^* = J_{u^*}$.The optimal value function J^* can be obtained by solving the following linear program :

$$\begin{aligned} &\min_{J \in \mathrm{R}^n} c^\top J \\ &\text{s.t. } J(s) \geq g_a(s) + \alpha \sum_{s'} p_a(s,s') J(s'), s \in S, a \in A. \end{aligned} \tag{1}$$

and the optimal policy can be obtained as

$$u^*(s) = \arg\max_{a \in A} \left(g_a(s) + \alpha \sum_{s'} p_a(s,s') J^*(s') \right) \tag{2}$$

We now define certain important quantities which will used in the rest of the discussion.

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 $(TJ)(s) = \max_{a \in A} \left(g_a(s) + \alpha \sum_{s'} p_a(s,s') J(s') \right)$.

(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 $\beta_\chi = \max_{s \in S} \frac{\max_{a \in A} \left(\alpha \sum_{s'} p_a(s,s') \chi(s') \right)}{\chi(s)}$.

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

[3]Such u^* exists and is well defined in the case of infinite horizon discounted reward MDP, for more details see [12].

(v) Function $\chi : S \to \mathrm{R}_+$ *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.*

When the MDP has a large number of states it is difficult to solve for J^*, using either the linear program (1) or other full state representation methods such as value iteration, policy iteration [2]).

III. Approximate Linear Programming

The approximate linear program (ALP) is obtained by making use of LFA in the LP, i.e., by adding the extra constraint $J = \Phi r$ in (1) with $\Phi \in \mathrm{R}^{n\times k}$ and introducing the new variables $r \in \mathrm{R}^k$. By substitution, this leads to

$$\begin{aligned} &\min_{r\in\mathrm{R}^k} c^\top \Phi r \\ &\text{s.t. } \Phi r \geq T\Phi r, \end{aligned} \tag{3}$$

where $J \geq TJ$ is a shorthand for the nd constraints in (1). Unless specified otherwise we use $\tilde{r}$ to denote any solution to the ALP and $\tilde{J} = \Phi\tilde{r}$ to denote the corresponding approximate value function. Further we assume the following to ensure the boundedness of the various linear programs (Eqs. (1), (3) and (5)) presented in this paper.

Assumption III.1. *(i)* $c = (c(i), i = 1, \ldots, n) \in \mathrm{R}^n$ *is a positive probability distribution, i.e.,* $c(i) > 0$ *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) $W \in \mathrm{R}_+^{nd\times m}$ *is a full rank* $nd \times m$ *matrix (where* $m \ll nd$*) and each of its column-sums equals one.*

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

It is straightforward to check that the vector $\mathbf{1}$ when viewed as an $S \to \mathrm{R}_+$ function is a Lyapunov function. Further, by Item ii, $\mathbf{1}$ is trivially present in the column span of Φ, hence,

Item iv is not limiting. In what follows we will always assume that Assumptions i–iv hold. When Item iv holds, it follows that for any $J \in \mathrm{R}^n$, $t > 0$, $s \in S$,

$$\begin{aligned}(T(J + t\psi))(s) &= \max_a g_a(s) + \alpha \sum_{s'} p_a(s, s')J(s') \\ &\quad + t\alpha \sum_{s'} p_a(s, s')\psi(s') \\ &\leq \max_a g_a(s) + \alpha \sum_{s'} p_a(s, s')J(s') \\ &\quad + t\beta_\psi \psi(s) \\ &= (TJ)(s) + \beta_\psi \, t \, \psi(s),\end{aligned}$$

or, in short,

$$T(J + t\psi) \leq TJ + \beta_\psi \, t \, \psi \qquad (J \in \mathrm{R}^n, \, t > 0) \, . \tag{4}$$

We will now recall two results due to de Farias and Van Roy that bound the error due to the introduction of the extra constraint in the ALP.

Theorem III.1 (Theorem 4.2 of [4])**.** *Let $\tilde{r}$ be the solution to the ALP in (3), $\tilde{J}_c = \Phi\tilde{r}_c$. Then, it holds that*

$$||J^* - \tilde{J}||_{1,c} \leq \frac{2c^\top \psi}{1 - \beta_\psi} \min_r ||J^* - \Phi r||_{\infty,\psi} \, .$$

The next result the loss in performance of a policy that is greedy w.r.t. $J_{\tilde{u}}$:

Theorem III.2 (Theorem 3.1 of [4])**.** *Let $\tilde{u}$ be the greedy policy with respect to the solution $\tilde{J}$ of the ALP. Then,*

$$||J^* - J_{\tilde{u}}||_{1,c} \leq \frac{1}{1 - \alpha} ||J^* - \tilde{J}||_{1,c'} \, ,$$

where $c' = (1 - \alpha)c^\top (I - \alpha P_{\tilde{u}})^{-1}$.

Theorems III.1 and III.2 together imply that the ALP addresses both the control and prediction problems. Please refer to [4] for a more detailed treatment of the ALP.

Note that the ALP is a linear program in k ($\ll n$) variables as opposed to the LP in (1) which has n variables. Nevertheless, the ALP has nd constraints (same as the LP) which is an issue when n is large and calls for constraint approximation/reduction techniques. One proposal in the literature to overcome this hurdle is to employ a procedure known as constraint sampling,

Cs: Column generation, Dantzig-Wolf decomposition?

wherein a subset of the original constraints of the ALP are sampled to formulate a *reduced linear program* (RLP). The perfromance analysis of the RLP can be found in [5]. The RLP has been shown to perform well in experiments [4, 5, 8] in various domains such as Tetris and in network of queues. An alternative approach to handle the constraints is employ function approximation in the dual variables of the ALP [3]. However, [3] does not provide theoretical guarntees for the loss in performance due to such an approximation.

IV. Generalized Reduced Linear Program

The Generalized Reduced Linear Program is given as:

$$\begin{aligned} &\min_{r\in\mathcal{N}} \; c^\top \Phi r, \\ &\text{s.t. } \; W^\top E\Phi r \geq W^\top H\Phi r, \end{aligned} \tag{5}$$

where $\mathcal{N}$ is an additional constraint set (see IV.1) to ensure the boundedness of the solution. We denote the solution to the GRLP by $\hat{r}$ and the approximate value functions are denoted by $\hat{J} = \Phi\hat{r}$.

Assumption IV.1. *$\mathcal{N} \subset \mathrm{R}^k$ is compact and $\tilde{r} \in \mathcal{N} \subset \mathrm{R}^k$.*

In what follows, we build the analytical framework to characterize the performance of the GRLP and then discuss the implications of our results.

V. Error Analysis

The least upper bound (LUB) projection operator $\Gamma : \mathrm{R}^n \to \mathrm{R}^n$ is defined as below:

Definition V.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*

$$\begin{aligned} &\min_{r\in\mathcal{N}'} \; c^\top \Phi r \,, \\ &\textit{s.t. } \; \Phi r \geq TJ. \end{aligned} \tag{6}$$

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

$$(\Gamma J)(i) \doteq (\Phi r_{e_i,J})(i), \quad i = 1, \ldots, n \,. \tag{7}$$

The approximate least upper bound (ALUB) projection operator $\hat{\Gamma} : \mathrm{R}^n \to \mathrm{R}^n$ is defined as below:

Definition V.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*

$$\begin{aligned} &\min_{r \in \mathcal{N}'} \ c^\top \Phi r \,, \\ &\text{s.t. } \ W^\top E \Phi r \geq W^\top H J. \end{aligned} \tag{8}$$

Then, for $J \in \mathrm{R}^n$ define $\hat{\Gamma} J$ as

$$(\hat{\Gamma} J)(i) \doteq (\Phi r'_{e_i,J})(i), \ \ i = 1, \ldots, n \,. \tag{9}$$

Assumption V.1. *The set $\mathcal{N}'$ is such that $\mathcal{N}' = \mathcal{N} + t r_0$ for any $t \in \mathrm{R}$, where $r_0 \in \mathrm{R}^k$ such that $\Phi r_0 = \psi$ and $\mathcal{N}$ is as in (5) satisfying Assumption IV.1.*

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

$$\begin{aligned} &\min\{c^\top A x : A x \geq b + b_0, x \in \mathcal{N}'\} \\ &= \min\{c^\top A y : A y \geq b, y \in \mathcal{N}'\} + c^\top b_0. \end{aligned} \tag{10}$$

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

Lemma V.2. *We have*

$$||J^* - \Gamma J^*||_{\infty,\psi} \leq 2 ||J^* - \Phi r^*||_{\infty,\psi}. \tag{11}$$

Proof. Define $\varepsilon = \|J^* - \Phi r^*\|_{\infty,\psi}$ so that $J^* - \Phi r^* \leq \varepsilon \psi$. Then from Assumption V.1 it follows that $\Phi(r^* + \varepsilon r_0) = \Phi r^* + \varepsilon \psi \geq J^* = T J^*$. From the definition of Γ in (7) 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 V.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 $\mathcal{F}_J \doteq \{ \Phi r \ : W^\top E \Phi r \geq W^\top H J, r \in \mathcal{N} \}$. Choose any $i \in \{1, \ldots, n\}$. Since $J_1 \leq J_2$, from Item iii it follows that $W^\top H J_1 \leq W^\top H J_2$. Hence, $\mathcal{F}_{J_2} \subset \mathcal{F}_{J_1}$ and thus $(\hat{\Gamma} J_1)(i) \leq (\hat{\Gamma} J_2)(i)$. Since i was arbitrary, the result follows. □

Lemma V.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 $t > 0$,*

$$\hat{\Gamma}(J + t\psi) \leq \hat{\Gamma}J + \beta t\psi\,. \tag{12}$$

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

Proof. First, we show that for any $t \geq 0$, $J \in \mathrm{R}^n$, $\hat{\Gamma}(J - t\psi) \geq \hat{\Gamma}J - \beta t\psi$ also holds. To see this define $J' = J - t\psi$. Then, $J = J' + t\psi$, hence $\hat{\Gamma}J \leq \hat{\Gamma}J' + \beta t\psi$. 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 $J_2 - \varepsilon\psi \leq J_1 \leq J_2 + \varepsilon\psi$. By the monotonicity of $\hat{\Gamma}$, $\hat{\Gamma}(J_2 - \varepsilon\psi) \leq \hat{\Gamma}J_1 \leq \hat{\Gamma}(J_2 + \varepsilon\psi)$. Using (12), we get $\hat{\Gamma}J_2 - \beta\varepsilon\psi \leq \hat{\Gamma}J_1 \leq \hat{\Gamma}J_2 + \beta\varepsilon\psi$, i.e., $-\beta\varepsilon\psi \leq \hat{\Gamma}J_1 - \hat{\Gamma}J_2 \leq \beta\varepsilon\psi$, from which the result follows. □

Corollary V.5. *Let us note in passing that from this result and (4) it immediately follows that T is an $\|\cdot\|_{\infty,\psi}$-contraction with factor β_ψ.*

Lemma V.6. *The operator $\hat{\Gamma}$ satisfies (12) with $\beta = \beta_\psi$.*

Proof. By definition, for $1 \leq i \leq n$, $(\hat{\Gamma}(J + t\psi))(i) = \min\{e_i^\top \Phi r \,:\, W^\top E\Phi r \geq W^\top H(J + t\psi), r \in \mathcal{N}\}$. By Eq. (4), as $t > 0$, $H(J + t\psi) \leq HJ + t\beta\psi$ and hence $W^\top H(J + t\psi) \leq W^\top(HJ + t\beta\psi)$. Thus, $(\hat{\Gamma}(J + t\psi))(i) \leq \min\{e_i^\top \Phi r \,:\, W^\top E\Phi r \geq W^\top H(J + t\psi), r \in \mathcal{N}\}$. Now, using Lemma V.1 with $A = W^\top E\Phi$, $b = TJ$, $c = e_i$, $b_0 = \beta t\psi$ and $x_0 = t\beta r_0$, the statement follows as thanks to Item iv, $Ax_0 = b_0$ and thanks to Assumption V.1, $\mathcal{N} = \mathcal{N} + \alpha t r_0$. □

Theorem V.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 Lemma V.4 and Lemma V.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 the vector $\hat{J}$ dominates the vector $\hat{V}$:

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

Proof. Let $r_1, r_2, \ldots, r_n$ be solutions to the GRLP in (5) for $c = e_1, e_2, \ldots, e_n$, respectively, and define $V_0 \in \mathrm{R}^n$ by $V_0(i) = \min_{j=1,\ldots,n} (\Phi r_j)(i)$, $1 \leq i \leq n$.

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

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

Lemma V.9. *A vector $\hat{r} \in \mathrm{R}^k$ is a solution to GRLP (5) iff it solves the following program:*

$$\begin{aligned} &\min_{r \in \mathrm{R}^k} ||\Phi r - \hat{V}||_{1,c} \\ &\text{s.t. } W^\top E \Phi r \geq W^\top H \Phi r\,, \qquad r \in \mathcal{N}\,. \end{aligned} \tag{13}$$

Proof. We know from Lemma V.8 that $\hat{J} = \Phi \hat{r} \geq \hat{V}$, and thus the solutions to (5) do not change if we add the constraint $\Phi r \geq \hat{V}$. 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 V.10. *We have*

$$||J^* - \hat{V}||_{\infty,\psi} \leq \frac{2||J^* - \Phi r^*||_{\infty,\psi} + ||\Gamma J^* - \hat{\Gamma} J^*||_{\infty,\psi}}{1 - \beta_\psi}. \tag{14}$$

Proof. By the triangle inequality,

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

and so by reordering and another triangle inequality,

$$\|J^* - \hat{V}\|_{\infty,\psi} \tag{15}$$

$$\leq \frac{\|J^* - \hat{\Gamma} J^*\|_{\infty,\psi}}{1 - \beta_\psi} \tag{16}$$

$$\leq \frac{\|J^* - \Gamma J^*\|_{\infty,\psi} + \|\Gamma J^* - \hat{\Gamma} J^*\|_{\infty,\psi}}{1 - \beta_\psi}\,. \tag{17}$$

The proof is complete by using Lemma V.2 in (15). □

We now recall Lemma 5 from Section 4.2 of [4]. For this result, recall that $r_{\in}\mathrm{R}^k$ is the vector such that $\psi = \Phi r_0$ by assumption.

Lemma V.11. *For $r \in \mathrm{R}^k$ arbitrary vector, let r' be*

$$r' = r + ||J^* - \Phi r||_{\infty,\psi} \left(\frac{1+\beta_\psi}{1-\beta_\psi}\right) r_0. \tag{18}$$

Then r' is feasible for the ALP in (3).

Recall that $\hat{V}$ is the fixed point of $\hat{\Gamma}$ and $\hat{J} = \Phi\hat{r}$ is the solution to the GRLP (5).

Theorem V.12. *We have*

$$\begin{aligned} ||\hat{J} - \hat{V}||_{1,c} \le \frac{c^\top\psi}{1-\beta_\psi}(4||J^* - \Phi r^*||_{\infty,\psi} \\ + ||\Gamma J^* - \hat{\Gamma} J^*||_{\infty,\psi}). \end{aligned} \tag{19}$$

Proof. Let $\gamma = ||J^* - \Phi r^*||_{\infty,\psi}$. Then, by choosing r' as in Lemma V.11 we have

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

Since r' is also feasible for the GRLP in (5) we have

$$\begin{aligned} ||\hat{J} - \hat{V}||_{1,c} &\le ||\Phi r' - \hat{V}||_{1,c} \\ &= \sum_{s\in S} c(s)\psi(s)\frac{|\Phi r'(s) - \hat{V}(s)|}{\psi(s)} \\ &\le c^\top\psi||\Phi r' - \hat{V}||_{\infty,\psi} \\ &\le c^\top\psi(||\Phi r' - J^*||_{\infty,\psi} + ||J^* - \hat{V}||_{\infty,\psi}). \end{aligned} \tag{20}$$

The result follows from Lemma V.10. □

Theorem V.13 (Prediction error bound in $\|\cdot\|_{\infty,\psi}$)**.** *It holds that*

$$||J^* - \hat{J}||_{1,c} \leq \frac{c^\top \psi}{1-\beta_\psi}(6||J^* - \Phi r^*||_{\infty,\psi} + 2||\Gamma J^* - \hat{\Gamma} J^*||_{\infty,\psi}). \tag{21}$$

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 V.10 and Theorem V.12. □

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

Theorem V.14 (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,*

$$\|J^* - J_{\hat{u}}\|_{1,c} \leq 2\left(\frac{c^\top \psi}{1-\beta_\psi}\right)^2 \big(6||J^* - \Phi r^*||_{\infty,\psi} + 2||\Gamma J^* - \hat{\Gamma} J^*||_{\infty,\psi}\big). \tag{22}$$

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,

$$\begin{aligned} ||J_{\hat{u}} - \hat{J}||_{1,c} &= ||(I - \alpha P_{\hat{u}})^{-1}(T\hat{J} - \hat{J})||_{1,c} \\ &\leq c^\top (I - \alpha P_{\hat{u}})^{-1} |T\hat{J} - \hat{J}| \\ &\leq c^\top (I - \alpha P_{\hat{u}})^{-1} \psi ||T\hat{J} - \hat{J}||_{\infty,\psi} \\ &\leq \frac{c^\top \psi}{1-\beta_\psi} ||T\hat{J} - \hat{J}||_{\infty,\psi} \\ &\leq \frac{c^\top \psi}{1-\beta_\psi}(||T\hat{J} - TJ^*||_{\infty,\psi} + ||J^* - \hat{J}||_{\infty,\psi}) \\ &\leq \frac{c^\top \psi}{1-\beta_\psi}(1+\beta_\psi)||J^* - \hat{J}||_{\infty,\psi}, \end{aligned} \tag{23}$$

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

$$\begin{aligned}
&||J^* - J_{\hat{u}}||_{1,c} \\
&\leq c^\top \psi ||J^* - \hat{J}||_{\infty,\psi} + c^\top \psi \frac{1+\beta_\psi}{1-\beta_\psi} ||J^* - \hat{J}||_{\infty,\psi} \\
&= \frac{2c^\top \psi}{1-\beta_\psi} ||J^* - \hat{J}||_{\infty,\psi}. \qquad (24)
\end{aligned}$$

The result now follows by substituting the bound on $||J^* - \hat{J}||_{\infty,\psi}$ from Theorem V.13. □

Note V.1. *By bounding* $||\Gamma J^* - \hat{\Gamma} J^*||_{\infty,\psi} = ||\Gamma J^* - J^* + J^* - \hat{\Gamma} J^*||_{\infty,\psi} \leq 2||J^* - \Phi r^*||_{\infty,\psi} + ||J^* - \hat{\Gamma} J^*||_{\infty,\psi}$ *(the inequality follows from Lemma V.2), we can loosen the bounds in Theorem V.13 and Theorem V.14 to*

$$\begin{aligned}
||J^* - \hat{J}||_{1,c} &\leq \frac{c^\top \psi}{1-\beta_\psi} (10||J^* - \Phi r^*||_{\infty,\psi} \\
&\quad + 2||J^* - \hat{\Gamma} J^*||_{\infty,\psi}). \qquad (25) \\
||J^* - J_{\hat{u}}||_{1,c} &\leq 2 \left(\frac{c^\top \psi}{1-\beta_\psi} \right)^2 \big(10||J^* - \Phi r^*||_{\infty,\psi} \\
&\quad + 2||J^* - \hat{\Gamma} J^*||_{\infty,\psi}\big). \qquad (26)
\end{aligned}$$

Here the term $||J^* - \hat{\Gamma} J^*||$ *in* (25) *and* (26) *captures the error due to the use of both* Φ *and* W*. Though,* (25) *and* (26) *might be loser bounds than* (21) *and* (22) *respectively, the advantage of this bound is that it captures the error due to function approximation as well as constraint reduction in a direct manner.*

The error term $||\Gamma J^* - \hat{\Gamma} J^*||_{\infty,\psi}$ gives new insights into constraint sampling.

Theorem V.15 (On Constraint Sampling)**.** *Let* $s \in S$ *be a state whose constraint was sampled. Then*

$$|\Gamma J^*(s) - \hat{\Gamma} J^*(s)| < |\Gamma J^*(s) - J^*(s)|. \qquad (27)$$

Proof. Let r_{e_s} and $\hat{r}_{e_s}$ be solutions to the linear programs in (6) and (8) respectively for $c = e_s$ and $J = J^*$. It is easy to note that r_{e_s} is feasible for the linear program in (8) for $c = e_s$ and J^*, and hence it follows that $(\Phi r_{e_s})(s) \geq (\Phi \hat{r}_{e_s})(s)$. However, since all the constraints with respect

to state s have been sampled we know that $(\Phi\hat{r}_{e_s})(s) \geq J^*$. The proof follows from noting that $(\Gamma J^*)(s) = (\Phi r_{e_s})(s)$ and $\hat{\Gamma} J^*(s) = (\Phi\hat{r}_{e_s})(s)$. □

VI. Discussion

In this section we discuss the implications and insights provided by the results presented in Theorems V.13 and V.14.

A. On Error Terms

- The error bounds in the main results (Theorems V.13 and V.14) contain two factors namely
  1) $\min_{r\in\mathrm{R}^k} ||J^* - \Phi r||_{\infty,\psi}$,
  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 [4].

  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.
- Unlike the result in [5] which holds only for a specific RLP formulated under ideal assumptions, our bounds hold for any GRLP and as a result for any given RLP. Another interesting feature of our result is that it holds with probability 1.
- A salient feature of the ALP formulation is the use of Lyapunov functions to control/shape the error across the states based on their relative importance. Since the error bounds are in a modified L_∞-norm, the GRLP framework retains this salient feature of the ALP.

The fact that both the prediction and control problems can be addressed by the GRLP makes it a complete ADP method, and by addressing the constraint approximation, the GRLP framework is an important addition to the theory of ALP.

B. On Constraint Reduction and Approximation

We claim the following based on the error bounds that we derived for the GRLP.

Claim 1) It is not always necessary to sample constraints according to the stationary distribution

of the optimal policy.

Claim 2) Constraint approximation is not only restricted to constraint sampling but also can be extended to include linear approximation of the constraints.

The following result (Theorem VI.1) supports Claim 1 in the above.

Main Result 3: On Constraint Sampling

The error term $||\Gamma J^* - \hat{\Gamma} J^*||_{\infty,\psi}$ gives new insights into constraint sampling.

Theorem VI.1. *Let $s \in S$ be a state whose constraint was sampled. Then*

$$|\Gamma J^*(s) - \hat{\Gamma} J^*(s)| < |\Gamma J^*(s) - J^*(s)|. \tag{28}$$

Proof. Let r_{e_s} and $\hat{r}_{e_s}$ be solutions to the linear programs in (6) and (8) respectively for $c = e_s$ and $J = J^*$. It is easy to note that r_{e_s} is feasible for the linear program in (8) for $c = e_s$ and J^*, and hence it follows that $(\Phi r_{e_s})(s) \geq (\Phi \hat{r}_{e_s})(s)$. However, since all the constraints with respect to state s have been sampled we know that $(\Phi \hat{r}_{e_s})(s) \geq J^*$. The proof follows from noting that $(\Gamma J^*)(s) = (\Phi r_{e_s})(s)$ and $\hat{\Gamma} J^*(s) = (\Phi \hat{r}_{e_s})(s)$. □

The expression in (28) in Theorem VI.1 says that the additional error $|\Gamma J^*(s) - \hat{\Gamma} J^*(s)|$ due to constraint sampling 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 (21) comes to our rescue to control the error due to those states that are not sampled. Thus the sampling distribution need not be the stationary distribution of the optimal policy as long as it samples the *important* states, an observation that might theoretically explain the empirical successes of the RLP [4, 8, 6].

To understand the implication of Claim 2 we need to look at the Lagrangian of the ALP and GRLP in (29) and (30) respectively, i.e.,

$$\tilde{L}(r, \lambda) = c^\top \Phi r + \lambda^\top (T\Phi r - \Phi r), \tag{29}$$

$$\hat{L}(r, q) = c^\top \Phi r + q^\top W^\top (T\Phi r - \Phi r). \tag{30}$$

The insight that the GRLP is a linear function approximation of the constraints (i.e., the Lagrangian multipliers) can be obtained by noting that $Wq \approx \lambda$ in (30). Note that 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). Further, W can be

interpreted as the feature matrix that approximates the Lagrange multipliers as $\lambda \approx Wq$, where $\lambda \in \mathrm{R}^{nd}, r \in \mathrm{R}^{m}$. One can show [7] that the optimal Lagrange multipliers are the discounted number of visits to the "state-action pairs" under the optimal policy u^*, i.e.,

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

where P_{u^*} is the probability transition matrix with respect to the optimal policy. 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 $\{P_{u^*}, P_{u^*}^2, \ldots\}$ hints at the kind of features that might be useful for the W matrix.

C. Reinforcement Learning

Reinforcement Learning (RL) algorithms are useful in scenarios where the system is available in the form of a simulator or only samples can be obtained via direct interaction. In particular, in the RL setting, the model parameters g and P are not known explicitly and the underlying MDP needs to be solved by using sample trajectories. In short, RL algorithms are sample trajectory based solution schemes for solving MDPs whose model information is not known. RL methods learn by filtering out the noisy sample via stochastic approximation and they also employ function approximation in order to handle MDPs with large number of states. Most RL algorithms are sample trajectory based extensions of ADP methods.

The RL extension of the ALP formulation has been applied to the optimal stopping problem in [3]. Function approximation is employed to approximate the square root of the Lagrange multipliers. However, since the approximation is not linear, convergence of the resulting RL algorithm cannot be guaranteed. Our results theoretically justify linear function approximation of the Lagrange multipliers, an immediate implication of which is that the RL extension of the ALP can be guaranteed to converge if the updates in [3] use LFA for the Lagrange multipliers instead of a non-linear approximator.

VII. Numerical Illustration

In this section, we show via an example in the domain of controlled queues [4] that the error term $||J^* - \Gamma J^*||_\infty$ indicates the quality of the constraint approximation. The queuing model used here is similar to the one in Section 5.2 of [4]. 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 [4]. For our experiments, we choose two contenders for the W-matrix and compare them with random positive matrix W_r. Our choices of the W matrix are as below.(**i**) W_c- matrix that corresponds to sampling according to c. This is justified by the insights obtained from Theorem VI.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 $\{P_{u^*}, P_{u^*}^2, \ldots\}$. Our choice of the W_a matrix to correspond to aggregation of near by 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 $i = 1, \ldots, m$,

$$\begin{aligned} W_a(i,j) = 1, \text{ for all } j \text{ s.t. } j = (i-1) \times \frac{n}{m} + k + (l-1) \times n, \\ k = 1, \ldots, \frac{n}{m}, l = 1, \ldots, d, \\ = 0, \text{ otherwise.} \end{aligned} \tag{31}$$

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 (31), W_c, W_i and W_r with $m = 50$. 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.

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.

A. Computational Complexity

It is known that the complexity of the linear program is polynomial in the length of its entries, i.e., the number of constraints times the number of variables [10, 1]. In the case of exact LP, both the number of variable as well as the number of constraints are of the order of the number of states. Thus, the complexity of obtaining the exact solution grows at a rate which is at least quadratic in the number of states. The entries of the GRLP on the other hand are of the order of $m \times k$, where the constants m and k are chosen to be much smaller than the number of states. We made use of the **SoPlex** solver and observed that solving the GRLPs (for $m = 50$, $k = 4$, $d = 4$) took a time of only about 0.03 seconds or lesser, the ALP with took about 0.7 seconds, while computing the exact solution took about 20 minutes. Also, it is important to note the sparsity of the W matrix in the case of state aggregation helps in formulating the constraints of the GRLP from the original constraints of the ALP without additional computational overhead.

Cs: $\times d$?

Cs: Sorry for the many questions, but did you also run ALP ($k = 4$) with no constraint aggregation?

VIII. CONCLUSION

The approximate linear programming (ALP) is a widely employed method to handle MDPs with large number of states. However, an important shortcoming of the ALP is that it has large number of constraints, which is tackled in practice by sampling a tractable number of constraints from the ALP to formulate and solve a reduced linear program (RLP). The RLP has been found to work well empirically in various domains ranging from queues to Tetris games, and performance guarantees are for a specific RLP formulated under idealized assumptions. Alternatively, function approximation in the dual variables of the ALP has been another approach that has been employed in literature [3, 7] to address the issue of large number of constraints. However [3, 7] do not provide theoretical characterization for the function approximation of the dual variables.

In this paper, we introduced a novel framework based on the generalized reduced linear program formulation to study constraint reduction. The constraints of the GRLP were obtained as positive linear combinations of the original ALP. We provided an error bound that relates the optimal value function to the solution of the GRLP. Our error bound contained two terms, one inherent to the ALP formulation and the other due to constraint reduction. We also made qualitative and quantitative observations about the nature of the error term that arose due to constraint reduction. Our analysis also revealed the fact that it is not always necessary to sample according to the stationary distribution of the optimal policy and, in fact, potentially several different constraint sampling/approximation strategies might work. In particular, we also theoretically justified linear function approximation of the constraints. We also discussed the results and provided a numerical example in the domain of controlled queues. To conclude, we observe that by providing a novel theoretical framework to study constraint approximation, this paper provides important results that add to the theory of ALP.

REFERENCES

[1] I. Adler and S. Cosares. A strongly polynomial algorithm for a special class of linear programs. *Operations Research*, 39(6):955–960, 1991. 9

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmont, MA, 4th edition, 2013. 2, 3

[3] V. S. Borkar, J. Pinto, and T. Prabhu. A new learning algorithm for optimal stopping. *Discrete Event Dynamic Systems*, 19(1):91–113, 2009. 2, 4, 8, 10

[4] D. P. de Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003. 2, 4, 6, 7, 8, 9

[5] 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. 2, 4, 7

[6] V. V. Desai, V. F. Farias, and C. C. Moallemi. A smoothed approximate linear program. In *NIPS*, pages 459–467, 2009. 2, 8

[7] 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. 8, 10

[8] 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. 2, 4, 8

[9] C. Guestrin, D. Koller, R. Parr, and S. Venkataraman. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intellegence Research*, 19:399–468, 2003. 2

[10] N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395, December 1984. 9

[11] J. Pazis and R. Parr. Non-parametric approximate linear programming for MDPs. In *AAAI*, 2011. 2

[12] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Programming*. John Wiley, New York, 1994. 2

[13] G. Taylor, M. Petrik, R. Parr, and S. Zilberstein. Feature selection using regularization in approximate linear programs for Markov decision processes. In *Proceedings of the Twenty-Seventh International Conference on Machine Learning*, Haifa, Israel, 2010. 2