MATHEMATICS OF OPERATIONS RESEARCH
Vol. 31, No. 3, August 2006, pp. 597–620
ISSN 0364-765X | EISSN 1526-5471 | 06 | 3103 | 0597

DOI 10.1287/moor.1060.0208

A Cost-Shaping Linear Program for Average-Cost Approximate Dynamic Programming with Performance Guarantees

Daniela Pucci de Farias
Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139,
pucci@mit.edu

Benjamin Van Roy
Departments of Management Science and Engineering and Electrical Engineering, Stanford University, Stanford, California 94305,
bvr@stanford.edu

We introduce a new algorithm based on linear programming for optimization of average-cost Markov decision processes (MDPs). The algorithm approximates the differential cost function of a perturbed MDP via a linear combination of basis functions. We establish a bound on the performance of the resulting policy that scales gracefully with the number of states without imposing the strong Lyapunov condition required by its counterpart in de Farias and Van Roy [de Farias, D. P., B. Van Roy. 2003. The linear programming approach to approximate dynamic programming. *Oper. Res.* **51**(6) 850–865]. We investigate implications of this result in the context of a queueing control problem.

Key words: approximate dynamic programming; linear programming; average cost
MSC2000 subject classification: Primary: 90C39, 90C40; secondary: 90C59
OR/MS subject classification: Primary: dynamic programming/optimal control, Markov infinite state; secondary: linear programming, large-scale systems
History: Received September 12, 2004; revised December 6, 2005.

1. Introduction. Over the past few years, there has been a growing interest in linear programming (LP) approaches to approximate dynamic programming (DP) (Adelman [1], de Farias and Van Roy [7, 9], Farias and Van Roy [11], Gordon [14], Guestrin [15], Guestrin et al. [16, 17], Hauskrecht and Kveton [19], Schuurmans and Patrascu [32], Schweitzer and Seidman [33], and Trick and Zin [unpublished manuscript, 1993; 34]). We refer to the methodology of LP-based approximate dynamic programming as *approximate linear programming* (ALP). The main idea in ALP is as follows. A classic result in dynamic programming is that Bellman's equation can be solved by solving a linear programming instead (D'Epenoux [10] and Manne [25]). ALP attempts to alleviate the curse of dimensionality by combining the LP formulation of Bellman's equation with approximation of the DP value function via a linear combination of preselected basis functions. The resulting LP is used for computing weights in the linear combination. A control policy that is "greedy" with respect to the resulting approximation is then used to make real-time decisions.

Empirically, ALP appears to generate effective control policies for high-dimensional dynamic programs (Adelman [1], de Farias and Van Roy [9], Farias and Van Roy [11], Guestrin et al. [17], and Schuurmans and Patrascu [32]). There is also evidence that, compared to other approximate dynamic programming approaches, ALP is competitive and its execution might in fact be orders of magnitude faster (Farias and Van Roy [11], Guestrin et al. [17], and Schuurmans and Patrascu [32]). At the same time, the strength of theoretical results about such algorithms has overtaken counterparts available for alternatives such as approximate value iteration, approximate policy iteration, and temporal-difference methods. As an example, a result in de Farias and Van Roy [9] implies that, for a discrete-time finite-state Markov decision process (MDP), if the span of the basis functions contains the constant function and comes within a distance of ϵ of the dynamic programming value function, then the approximation generated by a certain LP will come within a distance of $O(\epsilon)$. Here, the coefficient of the $O(\epsilon)$ term depends on the discount factor and the metric used for measuring distance but not on the choice of basis functions. On the other hand, the strongest results available for approximate value iteration and approximate policy iteration only promise $O(\epsilon)$ error under additional requirements on iterates generated in the course of executing the algorithms (Bertsekas and Tsitsiklis [3] and Munos [30]). In fact, it has been shown that even when $\epsilon = 0$, approximate value iteration can generate a diverging sequence of approximations (Boyan and Moore [5], Gordon [12, 13], and Tsitsiklis and Van Roy [35]).

We propose and analyze a new ALP formulation for approximating the dynamic programming solution. Previous ALP analysis found in the literature focuses on finite-state discrete-time MDPs with discounted-cost criterion. We consider average-cost problems in discrete and continuous time, involving countable state spaces. As a side benefit, the analysis leads to an apparently new condition for the solution of Bellman's equation via linear programming to be the optimal differential cost function, when the state space is countable. However, we consider the main contribution of the paper to be the derivation of an ALP formulation that is suitable for average-cost problems.

Challenges in average-cost ALP. The theoretical analysis of average-cost dynamic programming, especially in the case of infinite state spaces, is notoriously more involved than that of discounted-cost problems. In ALP the situation is no different. Previous analysis for discounted-cost problems establishes that, with an appropriate choice of algorithm parameters in ALP, the method exhibits certain desirable error guarantees. In particular, these guarantees suggest that the performance of the method will not degrade as it is applied to problems of increasing dimensions. However, attempts to extend the formulation and analysis to average-cost problems lead to at least two fundamental difficulties:

- As discussed in de Farias and Van Roy [9], a central concept in discounted-cost ALP is that of *state-relevance weights*. State-relevance weights are parameters in the ALP method that can be chosen to specify how errors in the approximation of the cost-to-go function over different system states should be emphasized. It is shown in de Farias and Van Roy [9] that state-relevance weights might have a first-order impact on the quality of the policy generated by ALP. State-relevance weights appear naturally when one extends the LP formulation of the discounted-cost Bellman's equation to approximate DP. However, the analogous extension in the average-cost setting, although seemingly intuitive and natural, does not include state-relevance weights and does not allow for specification of how to control the trade-off between approximation errors over different portions of the state space.
- Existing error analysis for discounted-cost ALP is based on the use of *Lyapunov functions* (de Farias and Van Roy [9]). The existence of a suitable Lyapunov function that is contained in the span of the basis functions used to approximate the DP solution is instrumental in ensuring that the ALP satisfies certain desirable properties. In particular, it ensures that the ALP is feasible and leads to approximation error bounds that scale gracefully with problem size. The main stumbling block in extending this analysis to average-cost problems is the condition that must be satisfied by the Lyapunov function. In the discounted-cost setting, the presence of the discount factor introduces some slack in the condition and allows for a suitable Lyapunov function to be derived. However, in the average-cost case the Lyapunov condition is overly restrictive.

In de Farias and Van Roy [8], we proposed a two-step ALP formulation for average-cost problems. The first step involves an LP for estimating the optimal average cost. The second step uses that estimate in another LP, which directly approximates the differential cost function. The formulation has the advantage of including state-relevance weights, which again can be used to control the trade-offs in approximating the differential cost function over different states. To extend the discounted-cost line of analysis, de Farias and Van Roy [8] introduce a slightly different definition of Lyapunov function. Under certain technical conditions (e.g., existence of a state that is recurrent under all policies), a Lyapunov function can be shown to exist. However, the resulting approximation error bound is unlikely to scale gracefully with problem size. In particular, the constants involved in the bound are expected to grow exponentially in the number of state variables of the system, which defeats the ALP objective of alleviating the curse of dimensionality. Moreover, even though a Lyapunov function can be shown to exist, finding such a function or ensuring that it is spanned by the basis functions could itself be a daunting task.

The cost-shaping LP and a performance bound. The aforementioned issues motivate the development of a new ALP formulation. We introduce a perturbed version of the MDP and add a *cost-shaping term* to the ALP. We derive a bound on the expected increase in average cost due to using the ALP policy in lieu of an optimal policy. The bound is stated in terms of the quality of the preselected basis functions. Specifically, the quality of a set of basis functions is measured by the best approximation error that can be attained if they are used to approximate the optimal differential cost function of the perturbed MDP. We show that the loss in performance related to the ALP policy is proportional in a certain sense to the best-possible approximation error. We consider the new formulation and associated performance bound to be the main contribution of this paper. Some of the most important aspects of the performance bound are discussed next.

Lyapunov-function-like bounds, without the Lyapunov condition. In the analysis of the new ALP, the cost-shaping term is shown to have a role analogous to the Lyapunov function. There are, however, two main differences. First, the cost-shaping term can be chosen arbitrarily and does not have to satisfy any condition. Hence, the Lyapunov function condition is relaxed. Furthermore, even when the cost-shaping term does satisfy the Lyapunov condition, the error and performance bounds that can be derived are strengthened. Specifically, they do not exhibit an undesirable dependence on the Lyapunov function that appears in the corresponding bounds obtained with the previous analysis (de Farias and Van Roy [9]).

Error vs. performance bounds. An important question about any approximate DP algorithm that approximates the solution to Bellman's equation is what guarantees can be established on the resulting approximation error. The analysis of ALP in de Farias and Van Roy [9] focuses primarily on that question, developing an error bound that establishes that, in terms of a certain choice of norm, the difference between the approximate and

exact cost-to-go functions is proportional to the best that can be achieved by the approximation architecture. Perhaps an even more important question is whether the policy that results from using the cost-to-go approximation exhibits good performance. This is partially addressed in de Farias and Van Roy [9] through a bound that relates the loss in performance due to using the suboptimal ALP policy to a certain norm of the approximation error. However, there is a mismatch between the norm that is used in the bound on the approximation error and the one that is required in the performance analysis. In this paper, we emphasize the performance aspect by phrasing our results in terms of the expected loss of performance due to using the suboptimal policy. The norm mismatch issue is captured through a constant present in the performance bound.

Relation to Bellman error minimization. An alternative approach for approximate DP aims at minimizing "Bellman error" (this idea was first suggested in Schweitzer and Seidman [33]). Methods proposed for this (e.g., Bertsekas [2] and Harmon [18]) involve stochastic steepest descent of a complex nonlinear function. There are no results indicating whether a global minimum will be reached or guaranteeing that a local minimum attained will exhibit desirable behavior. The LP we propose can be thought of as a method for minimizing a version of Bellman error. The important differences here are that our method involves solving a linear—rather than a nonlinear (and nonconvex)—program, and that performance guarantees can be made for the outcome.

Application to queueing control. We illustrate our approach through an application to service-rate control in a queueing system. The motivation is twofold. First, we derive performance bounds specialized to the queueing context. These are meant to illustrate how bounds on the approximation error and performance of the policy generated of the ALP can be derived for classes of problems. Hence, explicit a priori guarantees can be made about the behavior of ALP. In our analysis, we specify all algorithm parameters except for the selection of basis functions. We show that for all problems of service-rate control with the same structure as the one considered here, the approximation error and performance of the policy generated by ALP depends only on the quality of the choice of basis functions and certain problem parameters—most notably, the traffic intensity. To the best of our knowledge, no analogous guarantee exists for any other approximate DP method. The line of analysis presented here, which considers a class of problems rather than the performance of the method when applied to specific problem instances, is at this point unique to ALP.

The second purpose of the application of ALP to queueing control is to illustrate how the ALP method could be set up. In addition to basis functions, the proposed LP formulation requires selection of several algorithm parameters, all of which may have an impact on the approximation being generated and on the approximation error bound. At the current stage, the choice of parameters must be based on understanding of the problem at hand, and no automatic parameter selection algorithm exists. However, the error and performance analysis suggest some rules of thumb that might be useful in selecting these parameters. The selection of parameters in the class of queueing problems illustrates what factors should be taken in consideration and how these rules may be applied. We offer explicit values for all algorithm parameters, except for basis functions and reset probability, and discuss how the choices relate to the analysis in the paper.

Literature review. Our approximation method builds on the development of linear programming formulations for infinite-state average cost dynamic programming (see, e.g., Borkar [4] and Hernández-Lerma and Lasserre [21]). Earlier literature has also considered linear programming approximations for infinite-state problems in a spirit similar to what we present (Guestrin et al. [16], Hauskrecht and Kveton [19], and Trick and Zin [unpublished manuscript, 1993; 34]). However, none of these papers makes note of the absolute integrability requirement on basis functions, which is important for ensuring meaningful solutions. Furthermore, in this paper we offer a somewhat different algorithm together with performance bounds significantly stronger than those previously available. A detailed comparison with the results in de Farias and Van Roy [9] is given in the last section of this paper.

A fundamental aspect of our line of analysis is the use of weighted maximum norm. Such norms can be found in the literature on Markov chains and MDPs—in particular, weighted maximum norms appear in the analysis of stochastic shortest-path problems and algorithms for solving them (Bertsekas [2]). Weighted maximum norms are also used in the analysis of ergodicity of Markov chains (Kartashov [22, 23] and Meyn and Tweedie [28]).

Both theoretical and experimental work on value function approximation applied to queueing control can be found in the literature. Value and policy iteration are considered in Chen and Meyn [6], Henderson et al. [20], Meyn [26], and Meyn [27]. Approximate linear programming has also been applied to queueing control in Morrison and Kumar [29], where it is used for policy evaluation, and in Veatch [36], where special classes of basis functions are investigated.

Paper organization. The paper is organized as follows. The next section introduces formulations for discrete-time MDPs and introduces the notion of a perturbed MDP. Section 3 presents the LP approximation algorithm and error bound. The algorithm works with a perturbed MDP. Errors introduced by this perturbation

are studied in §4. In §5, we extend the analysis to continuous-time problems. The application to queueing control is discussed in §6. A closing section discusses relations to our prior work on LP approaches to approximate DP (de Farias and Van Roy [8, 9]).

2. Problem formulation. Consider an MDP evolving in discrete time with a countable state space $\mathcal{S}$ and a finite set of actions $\mathcal{A}$ available at each state. If an action $a \in \mathcal{A}$ is taken at a state $x \in \mathcal{S}$, the probability that the next state is y is denoted by $p_a(x, y)$. Under a stationary policy u: $\mathcal{S} \mapsto \mathcal{A}$, the state process follows a Markov chain with transition matrix $P_u \in \Re^{|\mathcal{S}|\times|\mathcal{S}|}$, in which each (x, y)th entry is $p_{u(x)}(x, y)$. A nonnegative cost $g(x, a) \geq 0$ is associated with each state-action pair (x, a). For shorthand, let $g_u(x) = g(x, u(x))$.

2.1. Average cost. The average cost λ_u under a stationary policy u is defined by

$$\lambda_u = \limsup_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \mathrm{E}\left[\sum_{t=0}^{\mathcal{T}-1} g_u(x_t) \middle| x_0 = x\right].$$

We assume that for each u, this value is independent of the initial state x. Let $\lambda^* = \inf_u \lambda_u$.

When the state space is finite, it can be shown that an average cost of λ^* is attained by a stationary policy, and further, that no nonstationary policy can attain lower average cost. These results do not necessarily hold when the state space is infinite. In particular, it might be the case that a nonstationary policy attains lower average cost than all stationary policies and/or no stationary policy attains average cost λ^*. For a treatment of issues arising with infinite state spaces and the average-cost criterion, see Puterman [31].

2.2. Discounted cost. An alternative objective is to minimize a discounted sum of expected future costs:

$$\mathrm{E}\left[\sum_{t=0}^{\infty} \alpha^t g_u(x_t) \middle| x_0 = x\right] = \sum_{t=0}^{\infty} \alpha^t (P_u^t g_u)(x),$$

where α is a discount factor in $(0, 1]$. In this context, it is useful to define a cost-to-go function

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

for each stationary policy u, as well as an optimal cost-to-go function

$$J^*(x) = \inf_u J_u(x).$$

Furthermore, we define a dynamic programming operator

$$(H_\alpha J)(x) = \min_{a\in\mathcal{A}} \left(g(x, a) + \alpha \sum_{y\in\mathcal{S}} p_{xy}(a) J(y) \right).$$

Nonnegativity of costs and finiteness of the action set imply the following:

(i) The optimal cost-to-go function solves Bellman's equation: $J^* = H_\alpha J^*$ (Bertsekas [2, Proposition 1.1, p. 137]).

(ii) A stationary policy u_α^* is optimal (among all policies—stationary or nonstationary) if and only if

$$u_\alpha^*(x) \in \arg\min_{a\in\mathcal{A}} \left(g(x, a) + \alpha \sum_{y\in\mathcal{S}} p_{xy}(a) J^*(y) \right).$$

(Bertsekas [2, Proposition 1.3, p. 143]).

Let u_α^* be an optimal stationary policy. Denote its average cost by λ_α^*. Let $c(x) > 0$ for all x, and let $\pi_{\alpha,u}(x) = (1-\alpha)\sum_{t=0}^{\infty} \alpha^t (c^T P_{u_\alpha^*}^t)(x)$, where c^T denotes the transpose of c. Consider the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & c^T J, \\ \text{subject to} \quad & H_\alpha J \geq J, \\ & \pi_{\alpha,u_\alpha^*}^T |J| < \infty. \end{aligned} \tag{1}$$

The following result motivates this optimization problem. Recall that we assume nonnegative costs and a finite action set.

THEOREM 2.1. *If u_α^* is an optimal stationary policy and $\pi_{\alpha,u_\alpha^*}^T J^* < \infty$, then J^* is the unique optimal solution to (1).*

2.3. Perturbation via restart. Under certain technical conditions,

$$\lim_{\alpha \uparrow 1}(1-\alpha)J_u(x) = \lambda_u$$

for all stationary policies u and states x. Hence, the discounted and average-cost objectives appear to become aligned as α approaches 1. This suggests that to deal with the difficulties arising in the average-cost problem, one might consider instead a discounted-cost formulation with large discount factor. The discounted-cost objective can be viewed as a perturbed version of the average-cost objective.

There is an alternative way to think about the perturbation that is equivalent to discounting costs but involves perturbing transition probabilities instead of the objective. The nature of the perturbation is influenced by two parameters: $\alpha \in (0, 1]$ and a distribution c over the state space. We refer to the new system as an (α, c)-perturbed MDP. It evolves similarly with the original MDP, except that at each time step, the state process "restarts" with probability $1-\alpha$; in this event, the next state is sampled randomly according to c. Hence, the perturbed MDP has the same state space, action space, and cost function as the original one, but the transition probability matrix under each stationary policy u is given by

$$P_{\alpha,u} = \alpha P_u + (1-\alpha)\mathbf{1}c^T.$$

We refer to $1-\alpha$ as the *restart probability* and c as the *restart distribution*.

In an (α, c)-perturbed MDP, the invariant distribution under each policy u is uniquely

$$\pi_{\alpha,u}^T = \lim_{t\to\infty} c^T P_{\alpha,u}^t = (1-\alpha)\sum_{t=0}^{\infty} \alpha^t c^T P_u^t. \tag{2}$$

Furthermore, the average cost of a stationary policy u is given by

$$\lambda_{\alpha,u} = \pi_{\alpha,u}^T g_u = \lim_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \sum_{t=0}^{\mathcal{T}-1} (P_{\alpha,u}^t g_u)(x) = (1-\alpha)\sum_{t=0}^{\infty} \alpha^t c^T P_u^t g_u = (1-\alpha)c^T J_u$$

for all x, where J_u is the cost-to-go function in the original MDP with a discount factor α. If $c(x) > 0$ for all x, a stationary policy u_α^* minimizes average cost in the perturbed MDP if and only if it minimizes discounted cost in the original MDP. Hence, optimization of the perturbed MDP with an average-cost objective is equivalent to optimization of the original MDP with a discounted objective.

Let $\lambda_\alpha^* = \min_u \lambda_{\alpha,u}$. For each stationary policy u, we define a differential cost function

$$h_{\alpha,u} = J_u - \frac{\lambda_{\alpha,u}}{1-\alpha}\mathbf{1}$$

and an optimal differential cost function

$$h_\alpha^* = J^* - \frac{\lambda_\alpha^*}{1-\alpha}\mathbf{1}.$$

Furthermore, we define dynamic programming operators

$$T_{\alpha,u}h = g_u + P_{\alpha,u}h \quad \text{and} \quad T_\alpha h = \min_u T_{\alpha,u}h,$$

where the minimization is taken pointwise. The following facts follow immediately from their analogs presented in the previous section:

(i) All pairs (λ, h) in the set $\{(\lambda_\alpha^*, h_\alpha^* + \kappa\mathbf{1}) | \kappa \in \Re\}$ are solutions to Bellman's equation, $h = T_\alpha h - \lambda\mathbf{1}$.

(ii) A stationary policy u_α^* is optimal (among all policies—stationary or nonstationary) if and only if

$$T_{\alpha,u_\alpha^*}h_\alpha^* = T_\alpha h_\alpha^*.$$

Let u_α^* be an optimal stationary policy. The following optimization problem is essentially equivalent to (1):

$$\begin{aligned} &\text{maximize} && \lambda, \\ &\text{subject to} && T_\alpha h - h - \lambda\mathbf{1} \geq 0, \\ & && \pi_{\alpha,u_\alpha^*}^T |h| < \infty. \end{aligned} \tag{3}$$

The following result is the analog to Theorem 2.1.

THEOREM 2.2. *If u_α^* is an optimal stationary policy and $\pi_{\alpha,u_\alpha^*}^T h_\alpha^* + \lambda^* < \infty$, then the set of optimal solutions to (3) is $\{(\lambda_\alpha^*, h_\alpha^* + \kappa\mathbf{1}) | \kappa \in \Re\}$.*

3. A linear programming approximation. Solving Bellman's equation for MDPs with infinite state spaces is infeasible in the absence of special structure. It requires computing and storing the differential cost function for each of the infinitely many states. In addition, even in the case of finite state spaces, solving Bellman's equation is often intractable due to the *curse of dimensionality*—the number of states grows exponentially in the number of state variables. We consider approximating h^*_α, the differential cost function of an (α, x)-perturbed MDP, using a linear combination $\sum_{k=1}^K r_k \phi_k$ of fixed basis functions $\phi_1, \ldots, \phi_K$: $\mathcal{S} \mapsto \Re$.

We now formulate and analyze an optimization problem for computing weights $r \in \Re^K$. Although the problem is not written as an LP, it is well known that it can be converted into one. For simplicity, we refer to it as an LP. It is useful to define a matrix $\Phi \in \Re^{|\mathcal{S}| \times K}$ so that our approximation to h^*_α can be written as Φr. Our problem takes as input several pieces of problem data:

(i) MDP parameters: $g(x, a)$ and $(P_u)_{xy}$ for all $x, y \in \mathcal{S}$, $a \in \mathcal{A}$, u: $\mathcal{S} \mapsto \mathcal{A}$.
(ii) Perturbation parameters: $\alpha \in [0, 1]$ and c: $\mathcal{S} \mapsto [0, 1]$ with $\sum_x c(x) = 1$.
(iii) Basis functions: $\Phi = [\phi_1 \ldots \phi_K] \in \Re^{|\mathcal{S}| \times K}$.
(iv) Slack function and penalty: ψ: $\mathcal{S} \mapsto [1, \infty)$ and $\eta > 0$.

We make the following assumption.

Assumption 3.1. *For some optimal policy* u^*_α,

$$\pi^T_{\alpha, u^*_\alpha} \left[J^* + \sum_{k=1}^K |\phi_k| + \psi \right] < \infty.$$

Recall from Theorem 2.2 that a sufficient condition for the solution of (3) to be a solution $(\lambda^*_\alpha, h^*_\alpha + \kappa \mathbf{1})$ to Bellman's equation is that h^*_α satisfy $\pi^T_{\alpha, u^*_\alpha} h^*_\alpha$. The previous assumption requires that the basis functions ϕ_k and slack function ϕ which are being used to approximate h^*_α, satisfy the same assumption.

The behavior of the new optimization problem as well as the performance bound we derive depends on the choice of parameters α, c, ψ, and η, in addition to the basis functions Φ. Note that there is potentially a very large number of parameters: α and η are scalars; however, both c and ψ are functions of the state. How to set such parameters automatically is still an open question. Appropriate choices can be obtained through a combination of problem-specific knowledge and insight derived from the error and performance analysis of ALP. We illustrate some key insights and rules of thumb in §6, in the context of queueing control.

We have defined all these terms except for the slack function and penalty, which we will explain after introducing the new ALP formulation. The ALP optimizes decision variables $r \in \Re^K$ and $s_1, s_2 \in \Re$ according to

$$\begin{aligned} \underset{r, s_1, s_2}{\text{minimize}} \quad & s_1 + \eta s_2, \\ \text{subject to} \quad & T_\alpha \Phi r - \Phi r + s_1 \mathbf{1} + s_2 \psi \geq 0, \\ & s_2 \geq 0. \end{aligned} \tag{4}$$

It is easy to see that (4) is feasible. It can also be shown that (4) is bounded, provided that η is large enough; the following lemma provides a sufficient condition.

Lemma 3.1. *If* $\eta \geq \pi^T_{\alpha, u^*_\alpha} \psi$, *then* (4) *is bounded.*

We denote an optimal solution by $(\tilde{r}, \tilde{s}_1, \tilde{s}_2)$. Although the first $|\mathcal{S}|$ constraints are nonlinear, each involves a minimization over actions and therefore can be decomposed into $|\mathcal{A}|$ constraints. This results in a total of $|\mathcal{S}| \times |\mathcal{A}| + 1$ constraints, which is unmanageable if the state space is large. We anticipate, however, that the solution to (4) can be approximated effectively through use of constraint-sampling techniques along the lines discussed in de Farias and Van Roy [7].

3.1. Relation to Bellman error minimization. Problem (4) can be viewed as a method for minimizing a form of Bellman error, as we now explain. Suppose that $s_2 = 0$. Then, minimization of s_1 corresponds to minimization of

$$\max_x ((\Phi r)(x) - \lambda^*_\alpha - (T_\alpha \Phi r)(x)),$$

which can be viewed as a measure of (one-sided) Bellman error. Measuring the maximum one-sided error over all states is problematic, however, when the state space is large. In the extreme case, when there is an infinite number of states and an unbounded cost function, such errors are typically infinite and therefore do not provide

a meaningful objective for optimization. This shortcoming is addressed by the slack term $s_2\psi$. To understand its role, consider constraining s_1 to be $-\lambda_\alpha^*$ and minimizing s_2. This corresponds to minimization of

$$\max_x \frac{(\Phi r)(x) - \lambda_\alpha^* - (T_\alpha \Phi r)(x)}{\psi(x)}.$$

This term can be viewed as a measure of weighted, one-sided Bellman error, with weights $1/\psi(x)$. One important factor that distinguishes our formulation from other approaches to Bellman error minimization (Bertsekas [2], Harmon et al. [18], and Schweitzer and Seidman [33]) is a theoretical performance guarantee, which we develop next.

3.2. A performance bound. For any r, let

$$u_{\alpha,r}(x) \in \arg\min_u \{g_u(x) + (P_{\alpha,u}\Phi r)(x)\}.$$

Let $\pi_{\alpha,r} = \pi_{\alpha,u_{\alpha,r}}$ and $\lambda_{\alpha,r} = \pi_{\alpha,r}^T g_{u_{\alpha,r}}$. Also define the weighted maximum norm

$$\|h\|_{\infty,\zeta} = \max_x \zeta(x)|h(x)|,$$

where $\zeta\colon \mathcal{S} \mapsto \Re^+$.

The following theorem establishes that the difference between the average cost $\lambda_{\alpha,\tilde{r}}$ associated with an optimal solution $(\tilde{r}, \tilde{s}_1, \tilde{s}_2)$ to problem (4) and the optimal average cost λ_α^* is proportional to the minimal error that can be attained given the choice of basis functions.

THEOREM 3.1. *If* $\eta \geq (2-\alpha)\pi_{\alpha,u_\alpha^*}^T\psi$, *then*

$$\lambda_{\alpha,\tilde{r}} - \lambda_\alpha^* \leq \frac{(1+\beta)\eta\max(\theta,1)}{1-\alpha} \min_{r\in\Re^K} \|h_\alpha^* - \Phi r\|_{\infty,1/\psi},$$

where

$$\beta = \max_u \|I - \alpha P_u\|_{\infty,1/\psi} \equiv \max_{u,h} \frac{\|(I-\alpha P_u)h\|_{\infty,1/\psi}}{\|h\|_{\infty,1/\psi}},$$

$$\theta = \frac{\pi_{\alpha,\tilde{r}}^T(T_\alpha\Phi\tilde{r} - \Phi\tilde{r} + \tilde{s}_1\mathbf{1} + \tilde{s}_2\psi)}{c^T(T_\alpha\Phi\tilde{r} - \Phi\tilde{r} + \tilde{s}_1\mathbf{1} + \tilde{s}_2\psi)}.$$

Theorem 3.1 offers insight as to how the algorithm parameters should be chosen. We postpone the discussion on the reset probability until §4. We show that $1-\alpha$ should be chosen to ensure that policies for the perturbed MDP offer similar performance when applied to the original MDP. This issue is related to mixing times of the policies of interest.

We discuss the choice of parameters c, ψ, and η in turn. The discussion also gives insight as to the meaning and behavior of the constants involved in the performance bound.

The choice of restart probability distribution c is related to constant θ. The bound suggests that the restart probability distribution c should be chosen to keep θ as small as possible. We can interpret the term $T_\alpha\Phi\tilde{r} - \Phi\tilde{r} + \tilde{s}_1\mathbf{1} + \tilde{s}_2\psi$, which appears in the definition of θ, as the Bellman error associated with the ALP approximation $\Phi\tilde{r}$, for the perturbed problem with cost shaping. This term is always nonnegative, because $(\Phi\tilde{r}, \tilde{s}_1, \tilde{s}_2)$ is a feasible solution to (4). Hence, θ can be interpreted as the ratio between two norms/expected values of the Bellman error, according to the reset distribution c and the stationary distribution $\pi_{\alpha,\tilde{r}}$. The ideal choice for c is $c = \pi_{\alpha,\tilde{r}}$, which gives the tightest bound in Theorem 3.1. This leads to a fixed-point problem because the solution $\tilde{r}$ depends on c itself. It is still an open question whether such a fixed point exists and can be found efficiently. However, having the reset distribution and the stationary distribution coincide is only a sufficient condition for making θ small; it is important only that the Bellman error as measured by each of these distributions be comparable. It is possible in some cases to use a priori knowledge about properties of stationary distributions in the application of interest to choose c so that θ remains within an acceptable range. In §6, we show how c can be chosen in the context of a queueing application so that θ is uniformly bounded on the size of the state space.

The bound suggests that the slack function ψ should be chosen so that the basis functions can offer a reasonable approximation error $\|h_\alpha^* - \Phi r\|_{\infty,1/\psi}$. At the same time, this choice affects the magnitudes of η and β. The theorem requires that the penalty η be at least $(2-\alpha)\pi_{\alpha,u_\alpha^*}^T\psi$. When dealing with specific classes of problems it

may be possible to select ψ so that the norm $\|h^*_\alpha - \Phi r\|_{\infty,1/\psi}$, as well as the terms β and $\pi^T_{\alpha,u^*_\alpha}\psi$, scale gracefully with the number of states and/or state variables. Often, a good compromise is having a function ψ that matches the growth of h^*_α. This ensures that $\|h^*_\alpha - \Phi r\|_{\infty,1/\psi}$ remains bounded and might lead to acceptable values of $\pi^T_{\alpha,u^*_\alpha}\psi$. This approach will be illustrated in §6. Note that, in general, if a certain state x has small stationary probability $\pi_{\alpha,u^*_\alpha}(x)$ under an optimal policy, then we can accordingly make $\psi(x)$ large. This captures the fact that such approximation errors over rarely visited portions of the state space are not as important and can be discounted by a large value of ψ. This is essential to the good scaling properties of ALP and the performance bound itself. Finally, it is worth pointing out that β, which also depends on the choice of ψ, is generally not of great concern. Specifically, it is easy to show that

$$\beta \le 1 + \alpha \max_{u,x} \frac{(P_u\psi)(x)}{\psi(x)}.$$

The second term in the right-hand side involves the ratio between the expected value of ψ at the next time step $(P_u\psi)(x)$ and its current value $\psi(x)$. We expect this term to be moderate for a wide range of problems and choices of ψ.

We conclude the discussion with the penalty term η. The main requirement is that $\eta \ge (2-\alpha)\pi^T_{\alpha,u^*_\alpha}\psi$. The performance bound in Theorem 3.1 is proportional to η; hence, it would be optimal to make η equal to the lower bound. Nevertheless, we do not know π_{α,u^*_α}, so that choosing η optimally or even ensuring that it satisfies the inequality may be difficult. One approach to selecting η is to perform a line search over possible values of η, solving (4) in each case, and choosing the value of η that results in the best-performing control policy. A simple line search algorithm solves (4) successively for $\eta = 1, 2, 4, 8, \ldots$, until the optimal solution is such that $\tilde{s}_2 = 0$. It is easy to show that (4) is unbounded for all $\eta < 1$ and that there is a finite $\bar{\eta} = \inf\{\eta | \tilde{s}_2 = 0\}$ such that for each $\eta \ge \bar{\eta}$, the solution is identical and $\tilde{s}_2 = 0$. This search process delivers a policy that is at least as good as a policy generated by (4) for some $\eta \in [(2-\alpha)\pi^T_{\alpha,u^*_\alpha}\psi, 2(2-\alpha)\pi^T_{\alpha,u^*_\alpha}\psi]$, and the performance bound of Theorem 3.1 would hold with η replaced by $2(2-\alpha)\pi^T_{\alpha,u^*_\alpha}\psi$.

4. The impact of perturbation. In Theorem 3.1 we compare the average costs $\lambda_{\alpha,\tilde{u}}$ and λ^*_α in the perturbed MDP. Under appropriate conditions, we expect that average costs of the perturbed MDP should converge to the average costs in the original policy, as α converges to one. Noting that the performance bound in Theorem 3.1 is proportional to $1/(1-\alpha)$, a relevant question is how large α has to be for the perturbed MDP to be an appropriate approximation. The answer is related to the notion of *mixing time*, which we define as follows.

Definition 4.1. The mixing time of policy u is given by

$$z_u = \inf\left\{ z: \left| \frac{1}{t}\sum_{t'=0}^{t-1} c^T P^{t'}_u g_u - \lambda_u \right| \le \frac{z}{t}, \ \forall\, t \right\}, \tag{5}$$

where $\inf \varnothing = +\infty$.

Condition (5) ensures that the average cost over a horizon of length $\frac{z_u}{\epsilon}$ is within ϵ of the long-run average cost for that policy, when the initial state is distributed according to c. It is interesting to compare this with existing definitions of mixing time. For instance, Kearns and Singh [24] define the *ϵ-mixing time* of a policy u to be the shortest time horizon after which the average cost is guaranteed to be within ϵ of the long-run average cost, at all times, for all initial states. Hence, their condition is weaker in the sense that it does not require that the difference between finite-horizon and infinite-horizon average costs be inversely proportional to the horizon, but it is stronger in the sense that it requires uniform mixing over all initial states. Indeed, we expect that in most problems the presence of the reset distribution in (5) is essential in ensuring that the mixing time will scale gracefully with problem size and lead to similarly graceful performance bounds.

When the state space is finite, condition (5) is always satisfied for some $z_u < \infty$. We expect that, in many practical contexts involving infinite state spaces, a suitable choice of c will also ensure that $z_u < \infty$ exists.

The following theorem gives a bound on the difference between the average costs of policy u in the original and perturbed MDPs as a function of its mixing time z_u.

Theorem 4.1. *For any stationary policy u, we have*

$$|\lambda_{\alpha,u} - \lambda_u| \le z_u(1-\alpha).$$

Based on Theorems 3.1 and 4.1, we have the following bound on the difference between the average cost of policy $\tilde{u}$ and the optimal average cost in the original MDP.

COROLLARY 4.1. *Let*

$$z^* = \lim_{\delta \downarrow 0} \inf\{z_u\colon \lambda_u \le \lambda^* + \delta\}.$$

Then,

$$\lambda_{\tilde{r}} - \lambda^* \le \frac{(1+\beta)\eta \max(\theta, 1)}{1-\alpha} \min_{r \in \Re^K} \|h_\alpha^* - \Phi r\|_{\infty, 1/\psi} + (1-\alpha)(z^* + z_{u_{\tilde{r}}}). \tag{6}$$

PROOF. Take an arbitrary $\delta > 0$ and any policy u such that $\lambda_u \le \lambda^* + \delta$. Then

$$\begin{aligned}
\lambda_{\tilde{r}} - \lambda^* &\le \lambda_{\tilde{r}} - \lambda_u + \delta \\
&\le \lambda_{\alpha, \tilde{r}} + |\lambda_{\tilde{r}} - \lambda_{\alpha, \tilde{r}}| - \lambda_{\alpha, u} + |\lambda_{\alpha, u} - \lambda_u| + \delta \\
&\le \lambda_{\alpha, \tilde{r}} - \lambda_\alpha^* + |\lambda_{\alpha, \tilde{r}} - \lambda_{\tilde{r}}| + |\lambda_u - \lambda_{\alpha, u}| + \delta \\
&\le \frac{(1+\beta)\eta \max(\theta, 1)}{1-\alpha} \min_{r \in \Re^K} \|h_\alpha^* - \Phi r\|_{\infty, 1/\psi} + (1-\alpha)(z_u + z_{u_{\tilde{r}}}) + \delta.
\end{aligned}$$

The third inequality follows from $\lambda_\alpha^* \le \lambda_{\alpha, u}$, and the fourth inequality follows from Theorems 3.1 and 4.1. Because δ is arbitrary, the result follows. □

Corollary 4.1 casts some light onto the question of how to choose the reset probability $1-\alpha$. As expected, a trade-off must be made. The larger the reset probability, the easier the perturbed problem becomes, whereas the relationship between average costs in the original and perturbed problems becomes weaker. In terms of the performance bound (6), the choice of α must strike a balance between two factors: the coefficient of $1/(1-\alpha)$ in the first term of (6) and the loss of

$$(1-\alpha)(z^* + z_{u_{\tilde{r}}}) \tag{7}$$

associated with the perturbation. Corollary 4.1 implies that in order for the absolute loss (7) to be less than or equal to ϵ, the coefficient $1/(1-\alpha)$ must be on the order of $z^* + z_{u_{\tilde{r}}}$. To simplify the discussion, let us assume that there is a policy u^* such that $\lambda_{u^*} = \lambda^*$. In this case, $z^* \le z_{u^*}$. It follows that $1/(1-\alpha)$ must be roughly proportional to the mixing times of policies $u_{\tilde{r}}$ and u^*.

For the performance bound (6) to scale gracefully with problem size, the mixing times $z_{u_{\tilde{r}}}$ and z^* must also scale gracefully. An important characteristic of the mixing time is that it is defined with respect to the restart distribution. Therefore, relatively fast mixing is required only on average over the possible initial states, rather than uniformly over all initial states in the system. This is important because if the state space is large, the mixing time starting from certain states could be very large. With a suitable choice of the reset distribution, this issue is bypassed.

Note that when $c \approx \pi_{\alpha, \tilde{r}} \approx \pi_{\tilde{r}}$, we have $\lambda_{\alpha, \tilde{r}} \approx \lambda_{\tilde{r}}$, and $1/(1-\alpha)$ must only be proportional to the mixing time of the optimal policy. An important open question is whether it is possible to design an approximate DP algorithm such that $c \approx \pi_{\alpha, \tilde{r}} \approx \pi_{\tilde{r}}$.

5. Continuous-time MDPs. The previous results extend naturally to continuous-time systems. Under a control policy $u\colon \mathcal{S} \mapsto \mathcal{A}$, the system dynamics in a continuous-time MDP are defined by an infinitesimal generator $A_u \in \Re^{|\mathcal{S}| \times |\mathcal{S}|}$, where each xth row of A_u depends only on $u(x)$. In particular, the probability mass function μ_t of x_t evolves according to

$$\dot{\mu}_t^T = -\mu_t^T A_u.$$

Let

$$\bar{A} = \max_{u, x, y} |A_u(x, y)|,$$

and assume that $\bar{A} < \infty$.

A *cost rate* $g(x, a)$ is associated with each state-action pair (x, a). As in the discrete case, we define $g_u(x) = g(x, u(x))$. Each stationary policy leads to an average cost λ_u given by

$$\lambda_u(x) = \lim_{\mathcal{T} \to \infty} \mathrm{E}\left[\frac{1}{T} \int_{t=0}^{\mathcal{T}} g_u(x_t)\, dt | x_0 = x\right],$$

and we assume that $\lambda_u(x) = \lambda_u$ for all x and some λ_u. A standard reduction from continuous to discrete time, known as uniformization (Puterman [31]), implies that results on the existence of solutions to Bellman's equations and optimality of the greedy policy with respect to the optimal cost-to-go function follow immediately from discrete-time counterparts.

We will consider an ALP formulation that fits the differential cost function in a perturbed version of the continuous-time MDP. For each policy u, we let

$$A_{\gamma,u} = A_u + \gamma(I - \mathbf{1}c^T),$$

where γ is a reset rate and c is a restart distribution. Associated with each policy u is a stationary distribution $\pi_{\gamma,u}$, which satisfies $\pi_{\gamma,u}A_{\gamma,u}^T = 0$, $\pi_{\gamma,u}^T\mathbf{1} = 1$. We denote by $\lambda_{\gamma,u} = \pi_{\gamma,u}^T g_u$ the average cost of policy u in the perturbed MDP. We define the discounted cost-to-go function

$$J_u(x) = \mathrm{E}\left[\int_{t=0}^{\infty} e^{-\gamma t} g_u(x_t)dt \,|\, x_0 = x\right]$$

and the optimal discounted cost-to-go function $J^* = \inf_u J_u$. Also define the differential cost functions

$$h_{\gamma,u} = J_u - \frac{\lambda_{\gamma,u}}{\gamma},$$

and

$$h_\gamma^* = J^* - \frac{\lambda_\gamma^*}{\gamma},$$

where $\lambda_\gamma^* = \inf_u \lambda_{\gamma,u}$.

It can be shown that a policy u_γ^* satisfies both $\lambda_{\gamma,u_\gamma^*} = \lambda_\gamma^*$ and $J_{u_\gamma^*} = J^*$. Moreover, all pairs (λ, h) in the set $\{(\lambda_{\gamma,u_\gamma^*}, h_\gamma^* + \kappa\mathbf{1}) | \kappa \in \Re\}$ satisfy the continuous-time version of Bellman's equation

$$F_\gamma h = \lambda\mathbf{1},$$

where the operator F_γ is defined by

$$F_\gamma h = \min_u\{g_u - A_{\gamma,u}h\},$$

and $h_\gamma^* + \kappa\mathbf{1}$, for any $\kappa \in \Re$, are the only solutions satisfying $\pi_{\gamma,u_\alpha^*}^T|h| < \infty$.

We consider the following optimization problem for generating an approximation to h_γ^*. It is entirely analogous to the discrete-time version (4).

$$\begin{aligned} \underset{r,s_1,s_2}{\text{minimize}} \quad & s_1 + \eta s_2, \\ \text{subject to} \quad & F_\gamma\Phi r + s_1\mathbf{1} + s_2\psi \geq 0, \\ & s_2 \geq 0. \end{aligned} \tag{8}$$

Let $(\tilde{r}, \tilde{s}_1, \tilde{s}_2)$ denote an optimal solution.

As in the discrete-time case, we make the following assumption on the basis functions and cost-shaping term.

Assumption 5.1. *For some optimal policy u_γ^*,*

$$\pi_{\gamma,u_\gamma^*}^T\left[J^* + \sum_{k=1}^K |\phi_k| + \psi\right] < \infty.$$

We have the following performance bound.

Theorem 5.1. *If $\eta \geq [(\bar{A}+2\gamma)/(\bar{A}+\gamma)]\pi_{\gamma,u_\gamma^*}^T\psi$, then*

$$\lambda_{\gamma,\tilde{r}} - \lambda_\gamma^* \leq \frac{(\bar{A}+\gamma)^2(1+\beta)\eta\max(\theta,1)}{\gamma}\min_{r\in\Re^K}\|h_\gamma^* - \Phi r\|_{\infty,1/\psi},$$

where

$$\beta = \max_u\left\|\frac{\gamma I - A_u}{\bar{A}+\gamma}\right\|_{\infty,1/\psi} \equiv \max_{u,h}\frac{\left\|\dfrac{\gamma I - A_u}{\bar{A}+\gamma}h\right\|_{\infty,1/\psi}}{\|h\|_{\infty,1/\psi}},$$

$$\theta = \frac{\pi_{\gamma,\tilde{r}}^T(F_\gamma\Phi\tilde{r} + \tilde{s}_1\mathbf{1} + \tilde{s}_2\psi)}{c^T(F_\gamma\Phi\tilde{r} + \tilde{s}_1\mathbf{1} + \tilde{s}_2\psi)}.$$

The significance of the bound and the interpretation of terms η, β, and θ are entirely analogous to that of the discrete-time case. We expect no substantial difference in the qualitative behavior of these quantities and, more generally, in the behavior of the continuous-time ALP.

6. Application: A queueing system. In this section, we consider application of ALP to a class of problems of service-rate control in a single-queue system. Our discussion is motivated by two main considerations. First, it illustrates how a combination of insights given by the performance analysis and exploitation of certain knowledge about the problem structure may lead to suitable choices for the ALP parameters. In particular, we state our results for specific choices of c, ψ, and η, leaving only the choice of basis functions Φ and reset rate γ generic. Second, the application illustrates how the performance analysis developed in the previous sections can be specialized to provide a priori guarantees for specific classes of problems.

We consider an infinite-buffer queue with Poisson arrivals at rate ℓ. There is a single server that can operate at m distinct service rates $\mu_1, \ldots, \mu_i$. A service rate of μ_i is associated with a usage cost of δ_i per unit time. The state of the system is the current queue length x_t, and at each time, a decision $a_t \in \{\mu_1, \ldots, \mu_m\}$ needs to be made on which service rate to choose. The total cost incurred at time t is $g(x_t, a_t) = x_t + \delta_{a_t}$.

A policy is a mapping from queue length to service-rate index. For each policy u: $\mathcal{Z}_+ \mapsto \{\mu_1, \ldots, \mu_m\}$, the infinitesimal generator is given by

$$(A_u)_{0y} = \begin{cases} \ell & \text{if } y = 0, \\ -\ell & \text{if } y = 1, \\ 0 & \text{otherwise;} \end{cases}$$

and for $x > 0$,

$$(A_u)_{xy} = \begin{cases} -u(x) & \text{if } y = x - 1, \\ \ell + u(x) & \text{if } y = x, \\ -\ell & \text{if } y = x + 1, \\ 0 & \text{otherwise.} \end{cases}$$

We assume that $\ell < \mu_1 < \mu_2 < \cdots < \mu_m$, i.e., the service rate of each server exceeds the arrival rate. Under this assumption, the process generated by each policy has a unique steady-state distribution π_u. For simplicity, we also assume $\bar{A} = \ell + \mu_m = 1$.

We define an auxiliary, fictitious service rate $\mu_0 = \sqrt{\ell\mu_1}$. Note that $\ell < \mu_0 < \mu_1$. We let $\bar{\delta} = \max_{1 \le i \le m} \delta_i = \delta_m$ and

$$\rho_i = \frac{\ell}{\mu_i}, \quad i = 0, \ldots, m.$$

We use continuous-time ALP to approximate the differential cost function. To do this, we must specify the algorithm parameters: $g, A_u, \gamma, \Phi, c, \psi, \eta$. The cost function g and infinitesimal generators A_u are as defined above. Let

$$\begin{aligned} \psi(x) &= x^2 + 1, \\ c(x) &= (1 - \rho_0)\rho_0^x, \\ \eta &= \frac{3 + 6\gamma}{(1 + \gamma)(1 - \rho_0)^2}. \end{aligned}$$

The motivation for these particular choices is as follows. It is well known that the differential cost function for this class of problems has asymptotic quadratic growth. Hence, the choice of quadratic ψ ensures that $\min_r \|h_\alpha^* - \Phi r\|_{\infty, 1/\psi}$ is not too large. It is interesting to note that the weighted maximum norm $\|\cdot\|_{\infty, 1/\psi}$ with quadratic ψ has also been used previously to analyze convergence of policy iteration for queueing control problems (Meyn [26]). It can also be shown that stationary distributions for this problem have exponential decay. The reset distribution c is chosen to exhibit the same structure, having exponential decay as well. Moreover, we choose the decay so that c decreases at a slower rate than any of the stationary distributions. This allows us to establish an a priori bound on θ. Finally, we exploit the fact that c decays more slowly than any stationary distribution to derive an upper bound on $(2 - \alpha)\pi_{\alpha, u_\alpha^*}^T \psi$. We choose η equal to this upper bound.

Note that we have not specified γ or Φ. The following theorems provide error bounds that depend on these two pieces of problem data. We make only the following assumption about Φ.

ASSUMPTION 6.1. $c^T|\phi_k| < \infty$, $k = 1, 2, \ldots, K$.

Theorem 6.1 establishes a bound on the loss in performance when the optimality criterion is the infinite-horizon discounted cost, with discount rate γ.

THEOREM 6.1. *For any* $m \geq 1$, $\ell > 0$, $\ell < \mu_i, i = 1, \ldots, m$, $\delta_1, \ldots, \delta_m \geq 0$, *and* $\gamma > 0$, *each bounded optimal solution* $\tilde{r}$ *to* (4) *satisfies*

$$c^T(J_{\tilde{r}} - J^*) \leq \frac{24(1+2\gamma)^3}{\mu_1(1-\rho_0)^4\gamma^2} \min_{r \in \Re^K} \|h^*_\gamma - \Phi r\|_{\infty, 1/\psi}.$$

Theorem 6.2 establishes a bound on the loss in performance when the optimality criterion is the infinite-horizon average cost.

THEOREM 6.2. *For any* $m \geq 1$, $\ell > 0$, $\ell < \mu_i$, $i = 1, \ldots, m$, $\delta_1, \ldots, \delta_m \geq 0$, *and* $\gamma > 0$, *each bounded optimal solution* $\tilde{r}$ *to* (4) *satisfies*

$$\lambda_{\tilde{r}} - \lambda^* \leq \frac{24(1+2\gamma)^3}{\mu_1(1-\rho_0)^4\gamma} \min_{r \in \Re^K} \|h^*_\gamma - \Phi r\|_{\infty, 1/\psi} + \gamma \frac{5 + 4\bar{\delta}}{2\mu_1^2(1-\rho_0)^4}.$$

7. Relation to prior work. In closing, it is worth discussing how our new algorithm and results relate to our prior work on LP approaches to approximate DP (de Farias and Van Roy [8, 9]). We will see that the slack function ψ serves a role that is equivalent to the role of Lyapunov functions in de Farias and Van Roy [9]. Consider the following optimization problem, which results from removing s_2 from (4):

$$\begin{aligned} &\text{minimize} \quad s_1, \\ &\text{subject to} \quad T_\alpha \Phi r - \Phi r + s_1 \mathbf{1} \geq 0. \end{aligned} \tag{9}$$

Let $(\hat{s}_1, \hat{r})$ be an optimal solution to (9). For any function $V: \mathcal{S} \mapsto \Re^+$, let

$$\beta_V = \alpha \| \max_u P_u V \|_{\infty, 1/V}.$$

We call V a *Lyapunov function* if $\beta_V < 1$. An analysis that parallels that of de Farias and Van Roy [9] would lead to the following result.

THEOREM 7.1. *If* $\beta_{\Phi v} < 1$ *and* $\Phi v' = \mathbf{1}$ *for some* $v, v' \in \Re^K$, *then,*

$$\lambda_{\alpha, \hat{r}} - \lambda^*_\alpha \leq \frac{2\max(\theta, 1) c^T \Phi v}{1 - \beta_{\Phi v}} \min_{r \in \Re^K} \|h^*_\alpha - \Phi r\|_{\infty, 1/\Phi v}.$$

A comparison of Theorems 3.1 and 7.1 reveals benefits afforded by the slack function. We consider the situation where $\psi = \Phi v$, which makes the bounds directly comparable. An immediate observation is that even though ψ and Φv play analogous roles in the bounds, ψ is not required to be a Lyapunov function. In this sense, Theorem 3.1 is stronger than Theorem 7.1. Moreover, if $\eta = \pi^T_{\alpha, u^*_\alpha} \psi$, we have

$$\frac{\eta}{1-\alpha} = c^T \sum_{t=0}^{\infty} \alpha^t P^t_{u^*_\alpha} \psi \leq \max_u c^T \sum_{t=0}^{\infty} \alpha^t P^t_u \Phi v \leq \frac{c^T \Phi v}{1 - \beta_V}.$$

The final term—which appears in the bound of Theorem 7.1—grows with the largest mixing time among all policies; whereas the first term—which appears in the bound of Theorem 3.1—depends only on the mixing time of an optimal policy. Hence, the coefficient of Theorem 3.1 should generally be much smaller than that of Theorem 7.1.

As discussed in de Farias and Van Roy [9], appropriate choice of c—there referred to as *state-relevance weights*—can be important for the error bound of Theorem 7.1 to scale well with the number of states. In de Farias and Van Roy [8], it is argued that some form of weighting of states in terms of a metric of relevance should continue to be important when considering average-cost problems. An LP-based algorithm is also presented in de Farias and Van Roy [8], but the results are far weaker than the ones we have presented in this paper, and we suspect that the LP-based algorithm of de Farias and Van Roy [8] will not scale well to high-dimensional problems.

In closing, we point out that the analysis in §6, which illustrates how suitable ALP parameters may be chosen for a class of problems in queueing control, is in no way exhaustive. In general, the choice of the various parameters involved in ALP, as well as the basis functions, should involve problem-specific analysis and experimentation. Part of the merit of the performance analysis, aside from providing guarantees for ALP, is that it suggests desirable properties for these parameters. It is an important open question whether automated or semiautomated methods for parameter selection could be developed.

Acknowledgments. This research was supported in part by the NSF under CAREER Grant ECS-9985229 and by the ONR under Grant MURI N00014-00-1-0637.

Appendix A. Proofs.

THEOREM 2.1. *If u_α^* is an optimal stationary policy and $\pi_{\alpha,u_\alpha^*}^T J^* < \infty$, then J^* is the unique optimal solution to (1).*

PROOF. Let J be an arbitrary feasible solution to (1). Note that

$$\begin{aligned}
&\sum_{x\in\mathcal{S}} c(x) \sum_{y\in\mathcal{S}} \sum_{z\in\mathcal{S}} \sum_{t=0}^{\infty} |\alpha^t P_{u_\alpha^*}^t(x,y)(I(y,z) - \alpha P_{u_\alpha^*}(y,z))J(z)| \\
&\quad \le \frac{1}{1-\alpha} \sum_{y\in\mathcal{S}} \sum_{z\in\mathcal{S}} \pi_{u_\alpha^*}(y)|(I(y,z) - \alpha P_{u_\alpha^*}(y,z))J(z)| \\
&\quad \le \frac{1}{1-\alpha} \sum_{y\in\mathcal{S}} \pi_{u_\alpha^*}(y)|J(y)| + \frac{\alpha}{1-\alpha} \sum_{y\in\mathcal{S}} \sum_{z\in\mathcal{S}} \pi_{u_\alpha^*}(y)|P_{u_\alpha^*}(y,z)J(z)| \\
&\quad = \frac{1}{1-\alpha} \pi_{u_\alpha^*}^T |J| + \frac{\alpha}{1-\alpha} \pi_{u_\alpha^*}^T P_{u_\alpha^*} |J| \\
&\quad \le \frac{2}{1-\alpha} \pi_{u_\alpha^*}^T |J| \\
&\quad < \infty.
\end{aligned}$$

Because $c(x) > 0$ for all x, it follows that, for each x,

$$\sum_{y\in\mathcal{S}} \sum_{z\in\mathcal{S}} \sum_{t=0}^{\infty} \left| \alpha^t P_{u_\alpha^*}^t(x,y)(I(y,z) - \alpha P_{u_\alpha^*}(y,z))J(z) \right| < \infty.$$

Therefore, we can use the Tonelli-Fubini theorem to arrive at

$$\sum_{t=0}^{\infty} \alpha^t P_{u_\alpha^*}^t \left((I - \alpha P_{u_\alpha^*})J\right) = \left(\sum_{t=0}^{\infty} \alpha^t P_{u_\alpha^*}^t (I - \alpha P_{u_\alpha^*}) \right) J = J.$$

The constraint $H_\alpha J \ge J$ implies that

$$(I - \alpha P_{u_\alpha^*})J \le g_{u_\alpha^*},$$

and it follows that

$$J = \sum_{t=0}^{\infty} \alpha^t P_{u_\alpha^*}^t (I - \alpha P_{u_\alpha^*})J \le \sum_{t=0}^{\infty} \alpha^t P_{u_\alpha^*}^t g_{u_\alpha^*} = J^*.$$

Recall that $J^* = H_\alpha J^*$ and $\pi_{\alpha,u_\alpha^*}^T J^* < \infty$. Hence, J^* is a feasible solution to (1). Furthermore, J^* is the unique optimum because $J \le J^*$ (for any feasible solution J) and $c(x) \ge 0$ for all x. □

THEOREM 2.2. *If u_α^* is an optimal stationary policy and $\pi_{\alpha,u_\alpha^*}^T J^* < \infty$, then the set of optimal solutions to (3) is $\{(\lambda_\alpha^*, h_\alpha^* + \kappa\mathbf{1}) | \kappa \in \Re\}$.*

PROOF. Consider (1), but with a change of variables $J = h + \lambda\mathbf{1}/(1-\alpha)$ where $c^T h = 0$. This gives

$$\begin{aligned}
\text{maximize} \quad & \lambda, \\
\text{subject to} \quad & H_\alpha\left(h + \frac{\lambda}{1-\alpha}\mathbf{1}\right) \ge h + \frac{\lambda}{1-\alpha}\mathbf{1}, \\
& c^T h = 0, \\
& \pi_{\alpha,u_\alpha^*}^T \left| h + \frac{\lambda}{1-\alpha}\mathbf{1} \right| < \infty.
\end{aligned}$$

From Theorem 2.1, the unique solution to this problem satisfies $h + \lambda/(1-\alpha) = J^*$, $c^T h = 0$, which yields $h = J^* - c^T J^* \mathbf{1} = h_\alpha^*$ and $\lambda = (1-\alpha)c^T J^* = \lambda_\alpha^*$. Some straightforward manipulations show that this is

equivalent to

$$\begin{aligned} \text{maximize} \quad & \lambda, \\ \text{subject to} \quad & T_\alpha h - h - \lambda \mathbf{1} \geq 0, \\ & c^T h = 0, \\ & \pi^T_{\alpha, u^*_\alpha} |h| < \infty. \end{aligned}$$

We conclude that the unique solution to this optimization problem is $(\lambda^*_\alpha, h^*_\alpha)$. Because $T_\alpha(h + \kappa \mathbf{1}) = Th + \kappa \mathbf{1}$ for any $\kappa \in \Re$, removing the constraint $c^T h = 0$ results in a set of optimal solutions $\{(\lambda^*_\alpha, h^*_\alpha + \kappa \mathbf{1}) | \kappa \in \Re\}$. □

LEMMA 3.1. *If $\eta \geq \pi^T_{\alpha, u^*_\alpha} \psi$, then (4) is bounded.*

PROOF. Let (r, s_1, s_2) be any feasible solution to (4). Consider a Markov process with costs $g_{u^*_\alpha} + s_2 \psi$ and transition probability matrix $P_{u^*_\alpha}$. Consider problem (3) for the α-perturbed version of this Markov process. Then (r, s_1) is a feasible solution because it satisfies

$$g_{u^*_\alpha} + s_2 \psi + P_{\alpha, u^*_\alpha} \Phi r \geq -s_1 \mathbf{1}.$$

Let J_ψ denote the cost-to-go function for this Markov process. Then we have

$$\begin{aligned} \pi^T_{\alpha, u^*_\alpha} J_\psi &= \pi^T_{\alpha, u^*_\alpha} \left[J^* + \sum_{t=0}^{\infty} \alpha^t P^t_{u^*_\alpha} \psi \right] \\ &< \infty, \end{aligned}$$

by Assumption 3.1. We conclude that Theorem 2.2 holds for this Markov process and

$$-s_1 \leq \lambda^*_\alpha + s_2 \pi^T_{\alpha, u^*_\alpha} \psi.$$

It follows that

$$\begin{aligned} s_1 + \eta s_2 &\geq -\lambda^*_\alpha - s_2 \pi^T_{\alpha, u^*_\alpha} \psi + s_2 \pi^T_{\alpha, u^*_\alpha} \psi \\ &= -\lambda^*_\alpha. \quad \square \end{aligned}$$

THEOREM 3.1. *If $\eta \geq (2 - \alpha) \pi^T_{\alpha, u^*_\alpha} \psi$ then*

$$\lambda_{\alpha, \tilde{r}} - \lambda^*_\alpha \leq \frac{(1 + \beta) \eta \max(\theta, 1)}{1 - \alpha} \min_{r \in \Re^K} \| h^*_\alpha - \Phi r \|_{\infty, 1/\psi},$$

where

$$\begin{aligned} \beta &= \max_u \| I - \alpha P_u \|_{\infty, 1/\psi} \equiv \max_u \frac{\| (I - \alpha P_u) h \|_{\infty, 1/\psi}}{\| h \|_{\infty, 1/\psi}}, \\ \theta &= \frac{\pi^T_{\alpha, \tilde{r}} (T_\alpha \Phi \tilde{r} - \Phi \tilde{r} + \tilde{s}_1 \mathbf{1} + \tilde{s}_2 \psi)}{c^T (T_\alpha \Phi \tilde{r} - \Phi \tilde{r} + \tilde{s}_1 \mathbf{1} + \tilde{s}_2 \psi)}. \end{aligned}$$

Before proving Theorem 3.1, we present the following lemma.

LEMMA A.1. *For every policy u, we have*

$$\pi_{\alpha, u} \geq (1 - \alpha) c.$$

PROOF. We have

$$\begin{aligned} \mathbf{1} \pi^{\mathcal{T}}_{\alpha, u} &= \lim_{\mathcal{T} \to \infty} \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} P^t_{\alpha, u} \\ &= \lim_{\mathcal{T} \to \infty} \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} P^{t-1}_{\alpha, u} (\alpha P_u + (1 - \alpha) \mathbf{1} c^T) \\ &\geq (1 - \alpha) \lim_{\mathcal{T} \to \infty} \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} P^{t-1}_{\alpha, u} \mathbf{1} c^T \\ &= (1 - \alpha) \mathbf{1} c^T. \end{aligned}$$

The lemma follows. □

We also have the following lemma.

LEMMA A.2. *Let* (r, s_1, s_2) *be any feasible solution to* (4). *Then*

$$\lambda_{\alpha,r} - \lambda_\alpha^* \le \frac{\max(\theta, 1)}{1-\alpha}(\lambda_\alpha^* + s_1 + \eta s_2).$$

PROOF. First note that

$$\begin{aligned}\sum_{x\in\mathcal{S}}\sum_{y\in\mathcal{S}} |\pi_{\alpha,u_\alpha^*}(x)P_{\alpha,u_\alpha^*}(x,y)\phi_k(y)r_k| &\le \sum_{x\in\mathcal{S}}\sum_{y\in\mathcal{S}} \pi_{\alpha,u_\alpha^*}(x)P_{\alpha,u_\alpha^*}(x,y)|\phi_k(y)||r_k| \\ &= \pi_{\alpha,u_\alpha^*}^T|\phi_k||r_k| \\ &< \infty,\end{aligned}$$

because by Assumption 3.1 $\pi_{\alpha,u_\alpha^*}^T|\phi_k| < \infty$. It follows from the Tonelli-Fubini theorem that

$$\begin{aligned}\pi_{\alpha,u_\alpha^*}^T(P_{\alpha,u_\alpha^*}\Phi r) &= (\pi_{\alpha,u_\alpha^*}^T P_{\alpha,u_\alpha^*})\Phi r \\ &= \pi_{\alpha,u_\alpha^*}^T \Phi r\end{aligned}$$

and $\pi_{\alpha,u_\alpha^*}^T(T_{\alpha,u_\alpha^*}\Phi r - \Phi r) = \lambda_\alpha^*$.

Since $T_\alpha \Phi r - \Phi r + s_1\mathbf{1} + s_2\psi \ge 0$, we have that

$$\begin{aligned}0 &\le \pi_{\alpha,u_\alpha^*}^T(T_\alpha\Phi r - \Phi r + s_1\mathbf{1} + s_2\psi) \\ &\le \pi_{\alpha,u_\alpha^*}^T(T_{\alpha,u_\alpha^*}\Phi r - \Phi r + s_1\mathbf{1} + s_2\psi) \\ &= \lambda_\alpha^* + s_1 + \pi_{\alpha,u_\alpha^*}^T\psi s_2.\end{aligned}$$

It follows that

$$\begin{aligned}\lambda_{\alpha,r} - \lambda_\alpha^* &= \pi_{\alpha,r}^T(g_{u_r} - \lambda_\alpha^*\mathbf{1}) \\ &= \pi_{\alpha,r}^T(T_\alpha\Phi r - \Phi r - \lambda_\alpha^*\mathbf{1}) \\ &\le \pi_{\alpha,r}^T(T_\alpha\Phi r - \Phi r + s_1\mathbf{1} + s_2\psi) + \pi_{\alpha,u_\alpha^*}^T\psi s_2 \\ &\le \theta c^T(T_\alpha\Phi r - \Phi r + s_1\mathbf{1} + s_2\psi) + \pi_{\alpha,u_\alpha^*}^T\psi s_2 \\ &\le \frac{\theta}{1-\alpha}\pi_{\alpha,u_\alpha^*}^T(T_\alpha\Phi r - \Phi r + s_1\mathbf{1} + s_2\psi) + \pi_{\alpha,u_\alpha^*}^T\psi s_2 \\ &\le \frac{\theta}{1-\alpha}\pi_{\alpha,u_\alpha^*}^T(T_{\alpha,u_\alpha^*}\Phi r - \Phi r + s_1\mathbf{1} + s_2\psi) + \pi_{\alpha,u_\alpha^*}^T\psi s_2 \\ &= \frac{\theta}{1-\alpha}(\lambda_\alpha^* + s_1 + \pi_{\alpha,u_\alpha^*}^T\psi s_2) + \pi_{\alpha,u_\alpha^*}^T\psi s_2 \\ &\le \frac{\max(\theta,1)}{1-\alpha}(\lambda_\alpha^* + s_1 + \eta s_2).\end{aligned}$$

The third inequality follows from Lemma A.1. □

PROOF OF THEOREM 3.1. Consider an optimal solution $(\tilde r, \tilde s_1, \tilde s_2)$ to (4). Because for any $r \in \Re^K$, the triplet $(r, -\lambda_\alpha^* + (1-\alpha)c^T(h_\alpha^* - \Phi r), \|T_\alpha\Phi r - \Phi r - \lambda_\alpha^*\mathbf{1} + (1-\alpha)c^T(h_\alpha^* - \Phi r)\mathbf{1}\|_{\infty,1/\psi})$ is a feasible solution to (4), we have

$$\begin{aligned}\lambda_{\alpha,\tilde r} - \lambda_\alpha^* &\le \frac{\max(\theta,1)}{1-\alpha}(\lambda_\alpha^* + \tilde s_1 + \eta\tilde s_2) \\ &\le \frac{\max(\theta,1)}{1-\alpha}\min_{r\in\Re^K}[(1-\alpha)c^T(h_\alpha^* - \Phi r) \\ &\quad + \eta\|T_\alpha\Phi r - \Phi r - \lambda_\alpha^*\mathbf{1} + (1-\alpha)c^T(h_\alpha^* - \Phi r)\mathbf{1}\|_{\infty,1/\psi}] \\ &\le \frac{\max(\theta,1)}{1-\alpha}\min_{r\in\Re^K}[(1-\alpha)c^T\psi\|h_\alpha^* - \Phi r\|_{\infty,1/\psi} \\ &\quad + \eta\|\min_u\{g_u + \alpha P_u\Phi r\} - \Phi r - \min_u\{g_u + \alpha P_u h_\alpha^*\} + h_\alpha^*\|_{\infty,1/\psi}]\end{aligned}$$

$$\leq \frac{\max(\theta,1)}{1-\alpha}[\pi^T_{\alpha,u^*_\alpha}\psi+\eta\beta]\min_{r\in\Re^K}\|h^*_\alpha-\Phi r\|_{\infty,1/\psi}$$
$$\leq \frac{(\beta+1)\max(\theta,1)\eta}{1-\alpha}\min_{r\in\Re^K}\|h^*_\alpha-\Phi r\|_{\infty,1/\psi}.$$

The first inequality follows from Lemma A.2. The second inequality follows from optimality of $(\tilde r,\tilde s_1,\tilde s_2)$. The fourth inequality follows from Lemma A.1 and the definition of β. □

THEOREM 4.1. *For any stationary policy u, we have*

$$|\lambda_{\alpha,u}-\lambda_u|\leq z_u(1-\alpha).$$

PROOF. We have

$$\begin{aligned}|\lambda_{\alpha,u}-\lambda_u| &= \left|(1-\alpha)\sum_{t=0}^{\infty}\alpha^t c^T P^t(g_u-\lambda_u\mathbf{1})\right|\\ &= (1-\alpha)\left|\sum_{\tau=0}^{\infty}(\alpha^\tau-\alpha^{\tau+1})\sum_{t=0}^{\tau}c^T P^t(g_u-\lambda_u\mathbf{1})\right|\\ &= (1-\alpha)^2\left|\sum_{\tau=0}^{\infty}\alpha^\tau(\tau+1)\left(\frac{1}{\tau+1}\sum_{t=0}^{\tau}c^T P^t g_u-\lambda_u\right)\right|\\ &\leq (1-\alpha)^2\sum_{\tau=0}^{\infty}\alpha^\tau(\tau+1)\frac{z_u}{\tau+1}\\ &= z_u(1-\alpha).\quad\square\end{aligned}$$

THEOREM 5.1. *If* $\eta\geq[(\bar A+2\gamma)/(\bar A+\gamma)]\pi^T_{\gamma,u^*_\gamma}\psi$, *then*

$$\lambda_{\gamma,\tilde r}-\lambda^*_\gamma\leq\frac{(\bar A+\gamma)^2(1+\beta)\eta\max(\theta,1)}{\gamma}\min_{r\in\Re^K}\|h^*_\gamma-\Phi r\|_{\infty,1/\psi},$$

where

$$\beta=\max_u\left\|\frac{\gamma I-A_u}{\bar A+\gamma}\right\|_{\infty,1/\psi}\equiv\max_u\frac{\left\|\dfrac{\gamma I-A_u}{\bar A+\gamma}h\right\|_{\infty,1/\psi}}{\|h\|_{\infty,1/\psi}},$$
$$\theta=\frac{\pi^T_{\gamma,\tilde r}(F_\gamma\Phi\tilde r+\tilde s_1\mathbf{1}+\tilde s_2\psi)}{c^T(F_\gamma\Phi\tilde r+\tilde s_1\mathbf{1}+\tilde s_2\psi)}.$$

PROOF. Let

$$P_u=I-\frac{A_u}{\bar A},$$
$$\alpha=\frac{\bar A}{\bar A+\gamma},$$
$$P_{\alpha,u}=\alpha P_u+(1-\alpha)\mathbf{1}c^T.$$

It is easy to show that

$$P_{\alpha,u}=I-\frac{A_{\gamma,u}}{\bar A+\gamma}.$$

The constraints of the continuous-time ALP problem (8) are equivalent to

$$g_u+P_{\alpha,u}[(\bar A+\gamma)\Phi r]+s_1\mathbf{1}+s_2\psi\geq[(\bar A+\gamma)\Phi r].$$

It follows that $(\tilde r,\tilde s_1,\tilde s_2)$ solves (8) iff $((\bar A+\gamma)\tilde r,\tilde s_1,\tilde s_2)$ solves (4). Moreover,

$$F_\gamma\Phi\tilde r=T\Phi(\bar A+\gamma)\tilde r-\Phi(\bar A+\gamma)\tilde r$$

and $u_{\gamma,\tilde r}=u_{\alpha,(\bar A+\gamma)\tilde r}$.

Consider the stationary distribution $\pi_{\gamma,u}$ for each policy u under the continuous-time perturbed MDP. Then

$$\pi_{\gamma,u}^T A_{\gamma,u} = 0 \Leftrightarrow,$$
$$\pi_{\gamma,u}^T (I - P_{\alpha,u}) = 0.$$

We conclude that $\pi_{\gamma,u} = \pi_{\alpha,u}$, and it follows that $\lambda_{\gamma,u} = \lambda_{\alpha,u}$ for all policies u. In particular, we may assume $u_\gamma^* = u_\alpha^*$. Also note that (λ, h) solve

$$F_\gamma h = \lambda \mathbf{1},$$

iff

$$T_\alpha(\bar{A} + \gamma)h = (\bar{A} + \gamma)h + \lambda \mathbf{1},$$

and we conclude that $h_\alpha^* = (\bar{A} + \gamma)h_\gamma^*$.

Let

$$\begin{aligned}
\beta &= \max_u \left\| \frac{\gamma I - A_u}{\bar{A} + \gamma} \right\|_{\infty, 1/\psi} \\
&= \max_u \|I - \alpha P_u\|_{\infty, 1/\psi}, \\
\theta &= \frac{\pi_{\gamma,\tilde{r}}^T (F_\gamma \Phi \tilde{r} + \tilde{s}_1 \mathbf{1} + \tilde{s}_2 \psi)}{c^T (F_\gamma \Phi \tilde{r} + \tilde{s}_1 \mathbf{1} + \tilde{s}_2 \psi)} \\
&= \frac{\pi_{\alpha,(\bar{A}+\alpha)\tilde{r}}^T (T_\alpha \Phi (\bar{A} + \alpha)\tilde{r} + \tilde{s}_1 \mathbf{1} + \tilde{s}_2 \psi)}{c^T (T_\alpha \Phi (\bar{A} + \alpha)\tilde{r} + \tilde{s}_1 \mathbf{1} + \tilde{s}_2 \psi)}.
\end{aligned}$$

Note that $\eta \geq [(\bar{A}+2\gamma)/(\bar{A}+\gamma)]\pi_{\gamma,u_\gamma^*}^T \psi$ is equivalent to $\eta \geq (2-\alpha)\pi_{\alpha,u_\alpha^*}^T \psi$. Hence, we can apply Theorem 3.1 to conclude that, if $\eta \geq [(\bar{A}+2\gamma)/(\bar{A}+\gamma)]\pi_{\gamma,u_\gamma^*}^T \psi$, we have

$$\begin{aligned}
\lambda_{\gamma,\tilde{r}} - \lambda_\gamma^* &= \lambda_{\alpha,(\bar{A}+\gamma)\tilde{r}} - \lambda_\alpha^* \\
&\leq \frac{(1+\beta)\eta \max(\theta, 1)}{1-\alpha} \min_{r \in \Re^K} \|h_\alpha^* - \Phi r\|_{\infty, 1/\psi} \\
&= \frac{(\bar{A} + \gamma)^2 (1+\beta)\eta \max(\theta, 1)}{\gamma} \min_{r \in \Re^K} \|h_\gamma^* - \Phi r\|_{\infty, 1/\psi}.
\end{aligned}$$

The second equality follows from $h_\alpha^* = (\bar{A} + \gamma)h_\gamma^*$ and $1 - \alpha = 1/(1+\gamma)$. □

THEOREM 6.1. *For any $m \geq 1$, $\ell > 0$, $\ell < \mu_i$, $i = 1, \ldots, m$, $\delta_1, \ldots, \delta_m \geq 0$, and $\gamma > 0$, each bounded optimal solution $\tilde{r}$ to (4) satisfies*

$$c^T (J_{\tilde{r}} - J^*) \leq \frac{24(1+2\gamma)^3}{\mu_1 (1-\rho_0)^4 \gamma^2} \min_{r \in \Re^K} \|h_\gamma^* - \Phi r\|_{\infty, 1/\psi}.$$

To prove Theorem 6.1, we start with a series of auxiliary lemmas.

For all u, let

$$P_u = I - A_u, \tag{10}$$
$$\alpha = \frac{1}{1+\gamma}, \tag{11}$$
$$P_{\alpha,u} = \alpha P_u + (1-\alpha)\mathbf{1}c^T = I - \frac{A_{\gamma,u}}{1+\gamma}. \tag{12}$$

Since $\bar{A} = \ell + \mu_m = 1$, P_u and $P_{\alpha,u}$ are transition probability matrices for discrete-time MDPs. Moreover,

$$\pi_{\gamma,u}^T = (1-\alpha)c^T \sum_{t=0}^{\infty} \alpha^t P_u^t \tag{13}$$

for all policies u.

LEMMA A.3. $\pi_{\gamma,u_\gamma^*}^T \psi < 3/(1-\rho_0)^2$.

PROOF. Let P_0 be the transition probability matrix associated with the (fictitious) policy $\hat{u}(x) = \mu_0$ for all $x > 0$, and let

$$J_0 = \sum_{t=0}^{\infty} \alpha^t P_0^t \psi.$$

It is easy to show that J_0 is well defined and $J_0 > \max_u \sum_{t=0}^{\infty} \alpha^t P_u^t$, for all u: $\mathcal{S} \mapsto \{\mu_1, \ldots, \mu_m\}$. It follows that

$$\begin{aligned}\pi_{\gamma, u_\gamma^*}^T \psi &= (1-\alpha) c^T \sum_{t=0}^{\infty} \alpha^t P_u^t \psi \\ &< (1-\alpha) c^T \sum_{t=0}^{\infty} \alpha^t P_0^t \psi \\ &= c^T \psi,\end{aligned}$$

where the last equality follows from the fact that c is the stationary distribution associated with P_0, i.e., $c^T P_0 = c^T$.

It follows from simple algebra that

$$\begin{aligned}c^T \psi &= \frac{\rho_0(\rho_0+1)}{(1-\rho_0)^2} + 1 \\ &< \frac{3}{(1-\rho_0)^2}. \quad \square\end{aligned}$$

LEMMA A.4. $\beta \le 3/(1+\gamma)$.

PROOF. We have

$$\begin{aligned}\beta &= \frac{1}{1+\gamma} \max_{u,h} \frac{\|(\gamma I - A_u)h\|_{\infty, 1/\psi}}{\|h\|_{\infty, 1/\psi}} \\ &= \frac{1}{1+\gamma} \max_{u,x} \frac{(|\gamma I - A_u|\psi)(x)}{\psi(x)} \\ &= \frac{1}{1+\gamma} \max_{x} \frac{(\ell + u(x) - \gamma)(x^2+1) + u(x)(x^2-2x+2) + \ell(x^2+2x+2)}{x^2+1} \\ &< \frac{1}{1+\gamma} \max_{x} \frac{(\ell + \mu_m)(x^2+1) + \mu_m(x^2-2x+2) + \ell(x^2+2x+2)}{x^2+1} \\ &= \frac{2}{1+\gamma} + \max_{x} \frac{2x(\ell-\mu_m)+1}{(1+\gamma)(x^2+1)} \\ &\le \frac{3}{1+\gamma},\end{aligned}$$

where we have used $\ell + \mu_m = 1$ and $\ell < \mu_m$. $\square$

LEMMA A.5. $\theta \le 2(1+\gamma)/[\mu_1(1-\rho_0)^2]$.

PROOF. Since $F_\gamma \Phi \tilde{r} + \tilde{s}_1 \mathbf{1} + \tilde{s}_2 \psi \ge 0$, we have

$$\begin{aligned}\theta &= \frac{\pi_{\gamma,\tilde{r}}^T (F_\gamma \Phi \tilde{r} + \tilde{s}_1 \mathbf{1} + \tilde{s}_2 \psi \ge 0)}{c^T (F_\gamma \Phi \tilde{r} + \tilde{s}_1 \mathbf{1} + \tilde{s}_2 \psi \ge 0)} \\ &\le \max_x \frac{\pi_{\gamma,\tilde{r}}(x)}{c(x)}.\end{aligned}$$

Recall that, for each u, $\pi_{\gamma,u}$ is the stationary distribution associated with the (discrete-time) transition probability matrix $P_{\alpha,u} = \alpha P_u + (1-\alpha)\mathbf{1}c^T$. Hence, $\pi_{\gamma,u}$ must satisfy the following balance equations $P(x_{t+1} > x, x_t \le x) = P(x_{t+1} \le x, x_t > x)$:

$$\alpha \pi_{\gamma,u}(x)\ell + (1-\alpha)\sum_{y \le x} \pi_{\gamma,u}(y) \sum_{y > x} c(y) = \alpha \pi_{\gamma,u}(x+1)u(x+1) + (1-\alpha)\sum_{y > x} \pi_{\gamma,u}(y) \sum_{y \le x} c(y).$$

Using $\sum_{y>x} c(y) = \rho_0^{x+1}$ and rearranging terms to isolate $\pi_{\gamma,u}(x+1)$, it follows that

$$\begin{aligned}\pi_{\gamma,u}(x+1) &= \frac{\ell}{u(x+1)} \pi_{\gamma,u}(x) + \frac{(1-\alpha)\rho_0^{x+1}}{\alpha u(x+1)} - \frac{1-\alpha}{\alpha u(x+1)} \left(1 - \sum_{y \le x} \pi_{\gamma,u}(y)\right) \\ &\le \frac{\ell}{\mu_1} \pi_{\gamma,u}(x) + \kappa \rho_0^{x+1} \\ &= \rho_0^2 \pi_{\gamma,u}(x) + \kappa \rho_0^{x+1},\end{aligned}$$

where $\kappa = (1-\alpha)/(\alpha\underline{u})$. Noting that $\pi_{\gamma,u}(0) \le 1$, we have for all x

$$\begin{aligned}\pi_{\gamma,u}(x) &\le \rho_0^{2x} + \kappa \sum_{y=1}^{x} \rho_0^y \rho_0^{2(x-y)} \\ &= \rho_0^{2x} + \kappa \frac{\rho_0^x - \rho_0^{2x}}{1-\rho_0},\end{aligned}$$

and

$$\begin{aligned}\frac{\pi_{\gamma,u}(x)}{c(x)} &\le \frac{\rho_0^x}{1-\rho_0} + \kappa \frac{1-\rho_0^x}{(1-\rho_0)^2} \\ &\le \frac{1}{1-\rho_0} + \frac{1-\alpha}{\alpha\mu_1(1-\rho_0)^2} \\ &= \frac{(1-\rho_0)\alpha\mu_1 + 1 - \alpha}{\alpha\mu_1(1-\rho_0)^2} \\ &< \frac{2}{\alpha\mu_1(1-\rho_0)^2} \\ &= \frac{2(1+\gamma)}{\mu_1(1-\rho_0)^2}. \quad \square\end{aligned}$$

Proof of Theorem 6.1. From Lemma A.3, we have

$$\eta = \frac{3+6\gamma}{(1+\gamma)(1-\rho_0)^2} \ge \frac{\bar{A}+2\gamma}{\bar{A}+\gamma} \pi_{\gamma,u_\gamma^*}^T \psi.$$

Therefore, Theorem 5.1 holds and we have

$$\begin{aligned}\lambda_{\gamma,\tilde{r}} - \lambda_\gamma^* &\le \frac{(1+\gamma)^2(1+\beta)\eta\theta}{\gamma} \min_r \|h_\gamma^* - \Phi r\|_{\infty,1/\psi} \\ &\le \frac{6(1+2\gamma)(4+\gamma)(1+\gamma)}{\mu_1(1-\rho_0)^4\gamma} \min_r \|h_\gamma^* - \Phi r\|_{\infty,1/\psi} \\ &< \frac{24(1+2\gamma)^3}{\alpha\mu_1(1-\rho_0)^4\gamma} \min_r \|h_\gamma^* - \Phi r\|_{\infty,1/\psi}. \qquad (14)\end{aligned}$$

The second inequality follows from Lemmas A.4 and A.5.

Finally, recall that for all u,

$$c^T J_u = \frac{\lambda_u}{\gamma}.$$

It follows that

$$c^T(J_{\tilde{r}} - J^*) = \frac{\lambda_{\gamma,\tilde{r}} - \lambda_\gamma^*}{\gamma}.$$

Applying (14), the theorem follows. $\square$

Theorem 6.2. *For any $m \ge 1$, $\ell > 0$, $\ell < \mu_i$, $i = 1,\dots,m$, $\delta_1,\dots,\delta_m \ge 0$, and $\gamma > 0$, each bounded optimal solution $\tilde{r}$ to (4) satisfies*

$$\lambda_{\tilde{r}} - \lambda^* \le \frac{24(1+2\gamma)^3}{\mu_1(1-\rho_0)^4\gamma} \min_{r\in\Re^K} \|h_\gamma^* - \Phi r\|_{\infty,1/\psi} + \gamma \frac{5+4\bar{\delta}}{2\mu_1^2(1-\rho_0)^4}.$$

As before, we will consider the discrete-time version of the MDP using the identities (10)–(13).

Lemma A.6. *Let $J(x) > 0$ be a nondecreasing function of x, and let $u_1(x) = \mu_1$ for all $x > 0$. Suppose that $\pi_{u_1}^T J < \infty$. Then*

$$\max_u \pi_u^T J = \pi_{u_1}^T J < c^T J.$$

Proof. First note that

$$\pi_u(x+1) = \frac{\ell}{u(x)} \pi_u(x),$$

and

$$\frac{\pi_u(x+1)}{\pi_{u_1}(x+1)} = \frac{u_1(x+1)}{u(x+1)}\frac{\pi_u(x)}{\pi_{u_1}(x)} \leq \frac{\pi_u(x)}{\pi_{u_1}(x)}.$$

It follows that, for all x such that $\pi_u(x) \geq \pi_{u_1}(x)$,

$$\sum_{y=0}^{x} \pi_u(x) \geq \sum_{y=0}^{x} \pi_{u_1}(x),$$

and for all x such that $\pi_u(x) < \pi_{u_1}(x)$,

$$\begin{aligned}\sum_{y=0}^{x} \pi_u(x) &= 1 - \sum_{y=x+1}^{\infty} \pi_u(x) \\ &\geq 1 - \sum_{y=x+1}^{\infty} \pi_{u_1}(x) \\ &= \sum_{y=0}^{x} \pi_{u_1}(x).\end{aligned}$$

Now

$$\begin{aligned}\pi_{u_1}^T J - \pi_u^T J &= J(0) + \sum_{x=1}^{\infty}(J(x) - J(x-1))\left(\sum_{y=x}^{\infty} \pi_{u_1}(y) - \pi_u(y)\right) \\ &\geq 0.\end{aligned}$$

The inequality follows from $J(x-1) < J(x),\ \forall x$ and $\sum_{y=x}^{\infty} \pi_{u_1}(y) - \pi_u(y) \geq 0$.

Finally,

$$\frac{\pi_{u_1}(x+1)}{c(x+1)} = \sqrt{\frac{\ell}{\mu_1}}\frac{\pi_{u_1}(x)}{c(x)} \leq \frac{\pi_{u_1}(x)}{c(x)}.$$

We conclude by an entirely analogous argument that $\pi_{u_1}^T J < c^T J$. □

Lemma A.7. *For each policy u, consider the discrete-time MDP with transition probabilities P_u, as given in (10). Let*

$$\begin{aligned}\mathcal{T} &= \inf\{t \geq 0\colon x_t = 0\}, \\ \tau &= \inf\{t > 0\colon x_t = 0\}.\end{aligned}$$

Then for all policies u,

(i) $E[\mathcal{T} | x_0 = x, a_t = u(x_t) \forall t] \leq x/(\mu_1 - \ell)$,
(ii) $E[\tau | x_0 = 0, a_t = u(x_t) \forall t] \leq \mu_1/(\mu_1 - \ell)$,
(iii) $E[\sum_{t=0}^{\mathcal{T}} x_t | x_0 = x, a_t = u(x_t) \forall t] \leq x^2/[2(\mu_1 - \ell)] + x/[2(\mu_1 - \ell)^2]$, *and*
(iv) $\lambda_u \leq \rho_1/(1-\rho_1) + \bar{\delta}$.

Proof. (i) Let $\hat{J}(x) = x/(\mu_1 - \ell)$. Then, $\hat{J}$ satisfies

$$\hat{J}(x) = \begin{cases} 1 + \ell\hat{J}(x+1) + \mu_1\hat{J}(x-1) + (1 - \ell - \mu_1)\hat{J}(x), & x > 0 \\ 0, & x = 0. \end{cases}$$

Let $\hat{u}(x) = \mu_1$ for all $x > 0$. Then $\pi_{\hat{u}}^T \hat{J} < \infty$, and it follows from Theorem 2.1 that

$$\hat{J} = \mathrm{E}[\mathcal{T} | x_0 = x, a_t = \hat{u}(x_t) \forall t].$$

We have, for any other policy u,

$$\hat{J}(x) \geq \begin{cases} 1 + \ell\hat{J}(x+1) + u(x)\hat{J}(x-1) + (1-\ell-u(x))\hat{J}(x), & x > 0 \\ 0, & x = 0. \end{cases}$$

It follows from standard dynamic programming arguments that

$$\mathrm{E}[\mathcal{T} \,|\, x_0 = x, a_t = u(x_t) \forall\, t] \leq \hat{J}(x).$$

(ii) For any policy u,

$$\begin{aligned} \mathrm{E}[\tau | x_0 = 0, a_t = u(x_t) \forall\, t] &= 1 + \ell \mathrm{E}[\mathcal{T} \,|\, x_0 = 1, a_t = u(x_t) \forall\, t] \\ &\leq 1 + \frac{\ell}{\mu_1 - \ell} \\ &= \frac{\mu_1}{\mu_1 - \ell}. \end{aligned}$$

(iii) Let

$$\hat{J}(x) = \frac{x^2}{2(\mu_1 - \ell)} + \frac{x(\mu_1 + \ell)}{2(\mu_1 - \ell)^2}.$$

Then $\hat{J}$ satisfies

$$\hat{J}(x) = \begin{cases} x + \ell\hat{J}(x+1) + \mu_1\hat{J}(x-1) + (1-\ell-\mu_1)\hat{J}(x), & x > 0 \\ 0, & x = 0. \end{cases}$$

Using an argument entirely analogous to that in Part 1 of this proof, we conclude that, for any policy u,

$$\mathrm{E}\left[\sum_{t=0}^{\mathcal{T}} x_t | x_0 = x, a_t = u(x_t) \forall\, t\right] \leq \frac{x^2}{2(\mu_1 - \ell)} + \frac{x(\mu_1 + \ell)}{2(\mu_1 - \ell)^2},$$

and the result follows from $\mu_1 + \ell \leq \mu_m + \ell = 1$.

(iv) We have

$$\begin{aligned} \max_u \lambda_u(x) &\leq \max_u \sum_x \pi_u(x) x + \bar{\delta} \\ &= \sum_x \pi_{u_1}(x) x + \bar{\delta} \\ &= \frac{\rho_1}{1 - \rho_1} + \bar{\delta}, \end{aligned}$$

where the first inequality follows from Lemma A.6. □

Lemma A.8. *For all policies u and all $\gamma > 0$,*

$$\lambda_{\gamma,u} - \lambda_u \leq \gamma \frac{5 + 4\bar{\delta}}{2\mu_1^2(1-\rho_0)^4}.$$

Proof. Let P_u and α be as given in (10) and (11). For each u, let $J_u = \sum_{t=0}^{\infty} \alpha^t P_u^t g_u$. Then,

$$\begin{aligned} \lambda_{\gamma,u} - \lambda_u &= \pi_{\gamma,u}^T g_u - \pi_u^T g_u \\ &= (1-\alpha)(c - \pi_u)^T \sum_{t=0}^{\infty} \alpha^t P_u^T g_u \\ &= (1-\alpha)(c - \pi_u)^T J_u. \end{aligned} \tag{15}$$

Consider the discrete-time MDP with transition probabilities P_u. Define the following sequence of stopping and interarrival times relative to when the system reaches state 0:

$$\begin{aligned} \mathcal{T}_{k+1} &= \inf\{t > \mathcal{T}_k \colon x_t = 0\}, \quad k = 0, 1, \ldots, \\ \mathcal{T}_0 &= -1, \\ \tau_k &= \mathcal{T}_{k+1} - \mathcal{T}_k, \quad k = 1, 2, \ldots. \end{aligned}$$

Then we can rewrite $J_u(x)$ as

$$\begin{aligned} J_u(x) &= \mathrm{E}\left[\sum_{t=0}^{\mathcal{T}_1}\alpha^t g_u(x_t)|x_0=x\right]+\mathrm{E}\left[\sum_{k=1}^{\infty}\sum_{t=\mathcal{T}_k+1}^{\mathcal{T}_{k+1}}\alpha^t g_u(x_t)|x_0=x\right] \\ &= \mathrm{E}\left[\sum_{t=0}^{\mathcal{T}_1}\alpha^t g_u(x_t)|x_0=x\right]+\sum_{k=0}^{\infty}\mathrm{E}\left[\alpha^{\mathcal{T}_1}|x_0=x\right]\mathrm{E}\left[\alpha^{\tau_1}\right]^k\mathrm{E}\left[\sum_{t=\mathcal{T}_1+1}^{\mathcal{T}_2}\alpha^t g_u(x_t)\right] \\ &= \mathrm{E}\left[\sum_{t=0}^{\mathcal{T}_1}\alpha^t g_u(x_t)|x_0=x\right]+\mathrm{E}\left[\alpha^{\mathcal{T}_1}|x_0=x\right]\frac{\mathrm{E}\left[\sum_{t=\mathcal{T}_1+1}^{\mathcal{T}_2}\alpha^t g_u(x_t)\right]}{1-\mathrm{E}\left[\alpha^{\tau_1}\right]}. \end{aligned} \tag{16}$$

We have

$$\begin{aligned} &\left|\sum_x (c(x)-\pi_u(x))\mathrm{E}\left[\sum_{t=0}^{\mathcal{T}_1}\alpha^t g_u(x_t)|x_0=x\right]\right| \\ &\quad\le \max\left[\mathrm{E}\left[\sum_{t=0}^{\mathcal{T}_1}\alpha^t g_u(x_t)|x_0\sim c\right],\mathrm{E}\left[\sum_{t=0}^{\mathcal{T}_1}\alpha^t g_u(x_t)|x_0\sim \pi_u\right]\right] \\ &\quad\le \max\left[\mathrm{E}\left[\sum_{t=0}^{\mathcal{T}_1} g_u(x_t)|x_0\sim c\right],\mathrm{E}\left[\sum_{t=0}^{\mathcal{T}_1} g_u(x_t)|x_0\sim \pi_u\right]\right] \\ &\quad= \mathrm{E}\left[\sum_{t=0}^{\mathcal{T}_1} g_u(x_t)|x_0\sim c\right] \\ &\quad\le \mathrm{E}\left[\sum_{t=0}^{\mathcal{T}_1}(x_t+\bar{\delta})|x_0\sim c\right] \\ &\quad\le \mathrm{E}\left[\frac{x_0^2}{2(\mu_1-\ell)}+\frac{x_0}{2(\mu_1-\ell)^2}+\frac{\bar{\delta}x_0}{\mu_1-\ell}|x_0\sim c\right] \\ &\quad= \frac{\rho_0(1+\rho_0)}{2(\mu_1-\ell)(1-\rho_0)^2}+\frac{\rho_0}{2(\mu_1-\ell)^2(1-\rho_0)}+\frac{\bar{\delta}\rho_0}{(\mu_1-\ell)(1-\rho_0)} \\ &\quad\le \frac{3+2\bar{\delta}}{2\mu_1^2(1-\rho_0)^3}. \end{aligned} \tag{17}$$

The equality follows from Lemma A.6, and the fourth inequality follows from Lemma A.7.

We also have

$$\begin{aligned} 0 &\le \sum_x(\pi_u(x)-c(x))\mathrm{E}[\alpha^{\mathcal{T}_1}|x_0=x]\frac{\mathrm{E}\left[\sum_{t=\mathcal{T}_1+1}^{\mathcal{T}_2}\alpha^t g_u(x_t)\right]}{1-\mathrm{E}[\alpha^{\tau_1}]} \\ &\le \sum_x(\pi_u(x)-c(x))\mathrm{E}[\alpha^{\mathcal{T}_1}|x_0=x]\frac{\mathrm{E}\left[\sum_{t=\mathcal{T}_1+1}^{\mathcal{T}_2} g_u(x_t)\right]}{1-\mathrm{E}[\alpha^{\tau_1}]} \\ &= \frac{\mathrm{E}[\alpha^{\mathcal{T}_1}|x_0\sim\pi_u]-\mathrm{E}[\alpha^{\mathcal{T}_1}|x_0\sim c]}{1-\mathrm{E}[\alpha^{\tau}]}\mathrm{E}[\tau]\lambda_u \\ &\le \frac{\mathrm{E}[1-\alpha^{\mathcal{T}_1}|x_0\sim c]}{1-\alpha}\mathrm{E}[\tau]\lambda_u \\ &\le \frac{\mathrm{E}[(1-\alpha)\mathcal{T}_1|x_0\sim c]}{1-\alpha}\mathrm{E}[\tau]\lambda_u \\ &\le \frac{\rho_0}{(\mu_1-\ell)(1-\rho_0)}\frac{\mu_1}{\mu_1-\ell}\left[\frac{\rho_1}{1-\rho_1}+\bar{\delta}\right] \\ &\le \frac{1+\bar{\delta}}{\mu_1(1-\rho_0)^4}. \end{aligned} \tag{18}$$

The first inequality follows from Lemma A.6 and the fact that $\mathrm{E}[\alpha^{\mathcal{T}_1}|x_0 = x]$ is decreasing in x. The first equality follows from $\mathrm{E}[\sum_{t=\mathcal{T}_1+1}^{\mathcal{T}_2} g_u(x_t)] = \mathrm{E}[\tau]\lambda_u$ (for a proof, see, e.g., Bertsekas [2]). The fourth inequality follows from

$$1 - \alpha^{\mathcal{T}_1} = \sum_{t=0}^{\mathcal{T}_1 - 1} (1-\alpha)\alpha^t \leq (1-\alpha)\mathcal{T}_1.$$

The fifth inequality follows from Lemma A.7.

From (15), (16), (17), and (18), we conclude that

$$\begin{aligned}
\lambda_{\gamma,u} - \lambda_u &= (1-\alpha)(c - \pi_u)^T J_u \\
&\leq (1-\alpha)\left|\sum_x (c(x) - \pi_u(x))\mathrm{E}\left[\sum_{t=0}^{\mathcal{T}_1} \alpha^t g_u(x_t)|x_0 = x\right]\right| \\
&\quad + (1-\alpha)\sum_x (\pi_u(x) - c(x))\mathrm{E}\left[\alpha^{\mathcal{T}_1}|x_0 = x\right]\frac{\mathrm{E}\left[\sum_{t=\mathcal{T}_1+1}^{\mathcal{T}_2} \alpha^t g_u(x_t)\right]}{1 - \mathrm{E}[\alpha^{\tau_1}]} \\
&\leq (1-\alpha)\frac{5 + 4\bar{\delta}}{2\mu_1^2(1-\rho_0)^4} \\
&= \frac{\gamma}{1+\gamma}\frac{5+4\bar{\delta}}{2\mu_1^2(1-\rho_0)^4} \\
&\leq \gamma\frac{5+4\bar{\delta}}{2\mu_1^2(1-\rho_0)^4}. \quad \square
\end{aligned}$$

Proof of Theorem 6.2. The result follows immediately from (14) and Lemma A.8.

Theorem 7.1. *If $\beta_{\Phi v} < 1$ and $\Phi v' = \mathbf{1}$ for some $v, v' \in \Re^K$, then,*

$$\lambda_{\alpha,\hat{r}} - \lambda_\alpha^* \leq \frac{2\max(\theta, 1)c^T\Phi v}{1 - \beta_{\Phi v}} \min_{r\in\Re^K} \|h_\alpha^* - \Phi r\|_{\infty, 1/\Phi v}.$$

Proof. It follows from straightforward algebraic manipulations that problem (9) can be rewritten as

$$\begin{aligned}
\text{minimize} \quad & -(1-\alpha)c^T\Phi\left[r - \left(c^T\Phi r + \frac{s_1}{1-\alpha}\right)v'\right], \\
\text{subject to} \quad & H_\alpha\Phi\left[r - \left(c^T\Phi r + \frac{s_1}{1-\alpha}\right)v'\right] - \Phi\left[r - \left(c^T\Phi r + \frac{s_1}{1-\alpha}\right)v'\right] \geq 0.
\end{aligned}$$

Denote an optimal solution to this problem by $(\hat{s}_1, \hat{r})$, and let $\bar{r}$ be an optimal solution to

$$\begin{aligned}
\text{minimize} \quad & -(1-\alpha)c^T\Phi r, \\
\text{subject to} \quad & H_\alpha\Phi r - \Phi r \geq 0.
\end{aligned} \tag{19}$$

It is clear that both problems have the same value, i.e.,

$$\hat{s}_1 = -(1-\alpha)c^T\Phi\bar{r}. \tag{20}$$

We now have

$$\begin{aligned}
\lambda_{\alpha,\hat{r}} - \lambda_\alpha^* &\leq \frac{\max(\theta,1)}{1-\alpha}(\lambda_\alpha^* + \hat{s}_1) \\
&= \max(\theta,1)c^T(J^* - \Phi\bar{r}) \\
&= \max(\theta,1)\|J^* - \Phi\bar{r}\|_{1,c} \\
&\leq \frac{2\max(\theta,1)c^T\Phi v}{1-\beta_{\Phi v}} \min_{r\in\Re^k} \|J^* - \Phi r\|_{\infty,1/\Phi v} \\
&\leq \frac{2\max(\theta,1)c^T\Phi v}{1-\beta_{\Phi v}} \min_{r\in\Re^k} \|h_\alpha * - \Phi r\|_{\infty,1/\Phi v}.
\end{aligned}$$

The first inequality follows from Lemma A.2 and the fact that $(\hat{r}, \hat{s}_1, 0)$ is a feasible solution to (4). The first equality follows from $\lambda_\alpha^* = (1-\alpha)c^T J^*$ and (20). The second equality follows from $J^* \geq \Phi\bar{r}$. The second

inequality follows from Theorem 4.2 in de Farias and Van Roy [9]—which can be shown to hold in the case of infinite state spaces as long as Parts 1 and 2 of Assumption 3.1 are satisfied—applied to problem (19). The third inequality follows from $h^*_\alpha = J^* - c^T J^* \mathbf{1}$ and the fact that $\mathbf{1}$ is in the span of Φ. □

References

[1] Adelman, D. 2004. A price-directed approach to stochastic inventory/routing. *Oper. Res.* **52**(4) 499–514.
[2] Bertsekas, D. P. 2001. *Dynamic Programming and Optimal Control*, 2nd ed. Athena Scientific, Belmont, MA.
[3] Bertsekas, D. P., J. N. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.
[4] Borkar, V. 2001. Convex analytic methods in Markov decision processes. E. Feinberg, A. Shwartz, eds. *Handbook of Markov Decision Processes: Methods and Applications*. Kluwer, Boston, MA.
[5] Boyan, J. A., A. W. Moore. 1995. Generalization in reinforcement learning: Safely approximating the value function. *Advances in Neural Information Processing Systems*, Vol. 7. MIT Press, Cambridge, MA.
[6] Chen, R.-R., S. Meyn. 1999. Value iteration and optimization of multiclass queueing networks. *Queueing Systems* **32** 65–97.
[7] de Farias, D. P., B. Van Roy. 2004. On constraint sampling in the linear programming approach to approximate dynamic programming. *Math. Oper. Res.* **29**(3) 462–478.
[8] de Farias, D. P., B. Van Roy. 2003. Approximate linear programming for average-cost dynamic programming. *Advances in Neural Information Processing Systems*, Vol. 15. MIT Press, Cambridge, MA.
[9] de Farias, D. P., B. Van Roy. 2003. The linear programming approach to approximate dynamic programming. *Oper. Res.* **51**(6) 850–865.
[10] D'Epenoux, F. 1963. A probabilistic production and inventory problem. *Management Sci.* **10**(1) 98–108.
[11] Farias, V. F., B. Van Roy. 2006. Tetris: A study of randomized constraint sampling. G. Calafiore, F. Dabbene, eds. *Probabilistic and Randomized Methods for Design Under Uncertainty*. Springer-Verlag, London, UK.
[12] Gordon, G. J. 1995. Stable function approximation in dynamic programming. Technical Report CMU-CS-95-103, Carnegie Mellon University, Pittsburgh, PA.
[13] Gordon, G. J. 1995. Stable function approximation in dynamic programming. *Machine Learning: Proc. Twelfth Internat. Conf.* (*ICML*), San Francisco, CA.
[14] Gordon, G. J. 1999. Approximate solutions to Markov decision processes. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.
[15] Guestrin, C. 2003. Planning under uncertainty in complex structured environments. Ph.D. thesis, Stanford University, Stanford, CA.
[16] Guestrin, C., M. Hauskrecht, B. Kveton. 2004. Solving factored MDPs with continuous and discrete variables. *Twentieth Conf. Uncertainty in Artificial Intelligence*, Banff, Alberta, Canada.
[17] Guestrin, C., D. Koller, R. Parr. 2003. Efficient solution algorithms for factored MDPs. *J. Artificial Intelligence Res.* **19** 399–468.
[18] Harmon, M. E., L. C. Baird, A. H. Klopf. 1995. Advantage updating applied to a differential game. *Advances in Neural Information Processing Systems*, Vol. 7. MIT Press, Cambridge, MA.
[19] Hauskrecht, M., B. Kveton. 2003. Linear program approximations to factored continuous-state Markov decision processes. *Advances in Neural Information Processing Systems*, Vol. 17. MIT Press, Cambridge, MA.
[20] Henderson, S. G., S. P. Meyn, V. B. Tadić. 2003. Performance evaluation and policy selection in multiclass networks. *Discrete Event Dynam. Systems: Theory Appl.* **13**(Special issue on learning, optimization and decision making (invited)) 149–189.
[21] Hernández-Lerma, O., J. B. Lasserre. 2001. The linear programming approach. E. Feinberg, A. Shwartz, eds. *Handbook of Markov Decision Processes: Methods and Applications*. Kluwer, Boston, MA.
[22] Kartashov, N. V. 1985. Criteria for uniform ergodicity and strong stability of Markov chains with a common phase space. *Theory Probab. Appl.* **30** 71–89.
[23] Kartashov, N. V. 1985. Inequalities in theorems of ergodicity and stability for Markov chains with a common phase space. *Theory Probab. Appl.* **30** 247–259.
[24] Kearns, M., S. Singh. 2002. Near-optimal reinforcement learning in polynomial time. *Machine Learning* **49**(2) 209–232.
[25] Manne, A. S. Linear programming and sequential decisions. *Management Sci.* **6**(3) 259–267.
[26] Meyn, S. P. 1997. The policy iteration algorithm for average reward Markov decision processes with general state space. *Trans. Automatic Control* **42**(12) 1663–1680.
[27] Meyn, S. P. 2005. Workload models for stochastic networks: Value functions and performance evaluation. *IEEE Trans. Automatic Control* **50**(8) 1106–1122.
[28] Meyn, S. P., R. Tweedie. 1993. *Markov Chains and Stochastic Stability*. Springer-Verlag.
[29] Morrison, J. R., P. R. Kumar. 1999. New linear program performance bounds for queueing networks. *J. Optim. Theory Appl.* **100**(3) 575–597.
[30] Munos, R. 2003. Error bounds for approximate policy iteration. *Machine Learning: Proc. Twentieth Internat. Conf.* (*ICML*). AAAI Press, Menlo Park, CA.
[31] Puterman, M. L. 1994. *Markov Decision Processes*. John Wiley & Sons, New York.
[32] Schuurmans, D., R. Patrascu. 2001. Direct value-approximation for factored MDPs. *Advances in Neural Information Processing Systems*, Vol. 14. MIT Press, Cambridge, MA.
[33] Schweitzer, P. J., A. Seidman. 1985. Generalized polynomial approximation in Markov decision processes. *J. Math. Anal. Appl.* **110** 568–582.
[34] Trick, M., S. Zin. 1997. Spline approximations to value functions: A linear programming approach. *Macroeconomic Dynam.* **1**(1) 255–277.
[35] Tsitsiklis, J. N., B. Van Roy. 1996. Feature–based methods for large-scale dynamic programming. *Machine Learning* **22** 59–94.
[36] Veatch, M. H. 2005. Approximate dynamic programming for networks: Fluid models and constraint reduction. Submitted for publication.