Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Dual Representations for Dynamic Programming

Tao Wang TRYSI@CS.ULABERTA.CA
Daniel Lizotte DLIZOTTE@CS.UALBERTA.CA
Michael Bowling BOWLING@CS.UALBERTA.CA
Dale Schuurmans DALE@CS.UALBERTA.CA
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8 Canada

Editor: Leslie Pack Kaelbling

Abstract

We propose to use a new dual approach to dynamic programming. The idea is to maintain an explicit representation of stationary distributions as opposed to value functions. A significant advantage of the dual approach is that it allows one to exploit well developed techniques for representing, approximating, and estimating probability distributions, without running the risks associated with divergent value function estimation. In this paper, we introduce novel dual representations and develop dual algorithms. Moreover, we show that dual algorithms remain stable in situations where standard value function estimation diverges. We also show that the dual view yields a viable alternative to standard value function based techniques and opens new avenues for solving sequential decision making problems.

Keywords: Sequential Decision Making, Dynamic Programming, Approximation

1. Introduction

Algorithms for dynamic programming (DP) and reinforcement learning (RL) are usually formulated in terms of *value functions*: representations of the long run expected value of a state or state-action pair (Sutton and Barto, 1998). The concept of value is so pervasive in DP and RL, in fact, that it is hard to imagine that a value function representation is not a necessary component of any solution approach. Yet, linear programming (LP) methods clearly demonstrate that the value function is not a necessary concept for solving DP/RL problems. In LP methods, value functions only correspond to the primal formulation of the problem, and do not appear at all in the dual. Rather, in the dual, value functions are replaced by the notion of state (or state-action) *visit distributions* (Puterman, 1994; Bertsekas, 1995; Bertsekas and Tsitsiklis, 1996). It is entirely possible to solve DP and RL problems in the dual representation, which offers an equivalent but different approach to solving DP/RL problems without any reference to value functions.

Despite the well known LP duality, dual representations have not been widely explored in DP and RL. In fact, they have only been anecdotally and partially treated in the RL literature (Dayan, 1993; Ng et al., 1999), and not in a manner that acknowledges any connection to LP duality. Nevertheless, as we will show, there exists a dual form for every

©2000 Tao Wang, Daniel Lizotte, Michael Bowling, and Dale Schuurmans.

standard value function algortihms including policy evaluation, policy iteration, Bellman iteration and for variants of these algorithms that use linear approximation.

In this paper, we offer a systematic investigation of dual solution techniques based on representing state visit and state-action visit distributions instead of value functions. Although many of our results show that the dual approach yields equivalent results to the primal approach in the tabular case —as one would expect—the dual approach has potential advantage over the primal in the approximate case as we will show below. The dual view offers a coherent and comprehensive perspective on optimal sequential decision making problems, just as the primal view, but offers new algorithmic insight and new opportunities for developing algorithms. In fact, there is the opportunity to develop a joint primal-dual view of sequential decision making under uncertainty, where combined algorithms might be able to exploit the benefits of both approaches in theoretically justified ways.

2. Preliminaries

We are concerned with the problem of optimal sequential decision making, and in particular, the problem of computing an optimal behavior strategy in a *Markov decision process* (MDP). An MDP is defined by a set of actions A, a set of states S, and

- an $|S||A|\times|S|$ *transition matrix* P, whose entries $P_{(sa,s')}$ specify the conditional probability of transitioning to state s' starting from state s and taking action a (hence P is nonnegative and *row* normalized), i.e., $P_{(sa,s')} = p(s' \mid s, a)$, where $p(s' \mid s, a) \geq 0$ and $\sum_{s'} p(s' \mid s, a) = 1 \ \forall \ s, a$.

- an $|S||A| \times 1$ *reward vector* $\mathbf{r}$, whose entries $\mathbf{r}_{(sa)}$ specify the reward obtained when taking action a in state s, i.e., $\mathbf{r}_{(sa)} = \mathrm{E}[r \mid s, a]$.

Here we focus on addressing the problem of computing an optimal behavior strategy for an MDP where the optimality criterion is maximizing the infinite horizon *discounted* reward $r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots = \sum_{t=1}^{\infty} \gamma^{t-1} r_t$ given a discount factor $0 \leq \gamma < 1$. It is known that an optimal behavior strategy can always be expressed by a stationary *policy*. In this paper, we will represent a stationary policy by an $|S||A| \times 1$ *vector* $\boldsymbol{\pi}$, whose entries $\boldsymbol{\pi}_{(sa)}$ specify the probability of taking action a in state s; that is $\sum_a \boldsymbol{\pi}_{(sa)} = 1$ for all s. Stationarity refers to the fact that the action selection probabilities do not change over time. In addition to stationarity, it is known that furthermore there always exists a *deterministic* policy that gives the optimal action in each state (i.e., simply a policy with probabilities of 0 or 1) (Bertsekas, 1995).

The main problem is to compute an optimal policy given either a complete specification of the environmental model P and $\mathbf{r}$ (the "*planning problem*"), or limited access to the environment through observed states and rewards and the ability to select actions to cause further state transitions (the "*learning problem*"). The planning problem is normally tackled by linear programming or dynamic programming methods, whereas the learning problem is solved by reinforcement learning methods.

3. Linear Programming

To establish the dual form of representation, we begin by briefly reviewing the LP approach for solving MDPs in the discounted reward case. Here we assume we are given the environmental variables P and $\mathbf{r}$, the discount factor γ, and the initial distribution over states, expressed by an $|S| \times 1$ vector $\boldsymbol{\mu}$. Then a standard LP for solving the planning problem can be expressed as (Puterman, 1994; Bertsekas, 1995; Bertsekas and Tsitsiklis, 1996)

$$\min_{\mathbf{v}} (1-\gamma)\boldsymbol{\mu}^\top \mathbf{v} \quad \text{subject to} \quad \mathbf{v}_{(s)} \geq \mathbf{r}_{(sa)} + \gamma P_{(sa,:)}\mathbf{v} \quad \forall s, a \tag{1}$$

where $\mathbf{v}$, an $|S| \times 1$ vector, is the state value function. By introducing an $|S| \times |S||A|$ marginalization matrix, which is built by placing $|S|$ row blocks of length $|A|$ in a block diagonal fashion, where each row block consists of all 1s. The primal LP (see Equation 1) can be rewritten as

$$\min_{\mathbf{v}} (1-\gamma)\boldsymbol{\mu}^\top \mathbf{v} \quad \text{subject to} \quad \Xi^\top \mathbf{v} \geq \mathbf{r} + \gamma P \mathbf{v} \tag{2}$$

That is, Ξ is constructed to simply ensure that the constraint $\Xi^\top \mathbf{v} \geq \mathbf{r} + \gamma P \mathbf{v}$ given in the above matrix form LP corresponds to the system of inequalities $\mathbf{v}_{(s)} \geq \mathbf{r}_{(sa)} + \gamma P_{(sa,:)}\mathbf{v}\ \forall s, a$ in the primal LP.

It is known that the optimal solution $\mathbf{v}^*$ to this LP corresponds to the *value function* for the optimal policy. In particular, given $\mathbf{v}^*$, the optimal policy can be recovered by

$$\begin{aligned} a^*(s) &= \arg\max_a \left(\mathbf{r}_{(sa)} + \gamma P_{(sa,:)}\mathbf{v}^*\right) \\ \boldsymbol{\pi}^*_{(sa)} &= \begin{cases} 1 & \text{if } a = a^*(s) \\ 0 & \text{if } a \neq a^*(s) \end{cases} \end{aligned}$$

Note that $\boldsymbol{\mu}$ and $(1-\gamma)$ behave as an arbitrary positive vector and positive constant in the LP above and do not affect the solution, provided $\boldsymbol{\mu} > 0$ and $\gamma < 1$ (de Farias and Van Roy, 2003). However, both play an important and non-arbitrary role in the dual LP below (as we will see) and we have chosen the objective in Equation 2 in a specific way to obtain the result below.

To derive our particular form of the dual LP, we first introduce a $|S||A| \times 1$ vector of Lagrange multipliers $\mathbf{d}$, and then form the Lagrangian of the standard LP (see Equation 2)

$$L(\mathbf{v}, \mathbf{d}) = (1-\gamma)\boldsymbol{\mu}^\top \mathbf{v} + \mathbf{d}^\top (\mathbf{r} + \gamma P \mathbf{v} - \Xi^\top \mathbf{v}), \ \ \mathbf{d} \geq 0$$

Next, taking the gradient of the Lagrangian with respect to $\mathbf{v}$ and setting the resulting vector equal to zero yields

$$\Xi \mathbf{d} = (1-\gamma)\boldsymbol{\mu} + \gamma P^\top \mathbf{d}$$

Substituting this constraint back into the Lagrangian eliminates the $\mathbf{v}$ variable and results in the dual LP.

$$\max_{\mathbf{d}} \mathbf{d}^\top \mathbf{r} \quad \text{subject to} \quad \mathbf{d} \geq 0, \ \ \Xi \mathbf{d} = (1-\gamma)\boldsymbol{\mu} + \gamma P^\top \mathbf{d} \tag{3}$$

Interestingly, the following lemma establishes that any feasible vector in the dual LP is guaranteed to be normalized, and therefore the solution $\mathbf{d}^*$ is always a joint *probability distribution* over state-action pairs.

Lemma 1 *If* $\mathbf{d}$ *satisfies the constraint in the dual LP (see Equation 3), then* $\mathbf{1}^\top \mathbf{d} = 1$.

Proof By definition of Ξ, we have $\mathbf{1}^\top \mathbf{d} = \mathbf{1}^\top \Xi \mathbf{d}$. Multiplying the constraint in Equation 3 by $\mathbf{1}^\top$ yields $\mathbf{1}^\top \Xi \mathbf{d} = (1-\gamma)\mathbf{1}^\top \boldsymbol{\mu} + \gamma \mathbf{1}^\top P^\top \mathbf{d}$. Since P is row normalized and $\boldsymbol{\mu}$ is a probability distribution, $\mathbf{1}^\top \Xi \mathbf{d} = (1-\gamma) + \gamma \mathbf{1}^\top \mathbf{d}$. Thus $\mathbf{1}^\top \mathbf{d} = (1-\gamma) + \gamma \mathbf{1}^\top \mathbf{d}$ holds. Since $0 \leq \gamma < 1$, $\mathbf{1}^\top \mathbf{d} = 1$. ■

By strong duality, we know that the optimal objective value of this dual LP equals the optimal objective value of the primal LP. Furthermore, given a solution to the dual $\mathbf{d}^*$, the optimal policy can be directly recovered by the much simpler transformation (Ross, 1997)

$$\boldsymbol{\pi}^*_{(sa)} \quad = \quad \frac{\mathbf{d}^*_{(sa)}}{\sum_a \mathbf{d}^*_{(sa)}} \tag{4}$$

A careful examination of the dual LP shows that the joint distribution $\mathbf{d}^*$ does *not* actually correspond to the stationary state-action visit distribution induced by $\boldsymbol{\pi}^*$ (unless $\gamma = 1$), but it does correspond to a distribution of discounted state-action visits beginning in the initial state distribution $\boldsymbol{\mu}$.

This dual LP formulation establishes that the optimal policy $\boldsymbol{\pi}^*$ for an MDP can be recovered without any direct reference whatsoever to the *value* function. Instead, one can work in the dual, and bypass value functions entirely, while working instead with *normalized* probability distributions over state-action pairs.

Before we present dual algorithms and their convergence analysis. We find it convenient to express a policy $\boldsymbol{\pi}$ as an $|S|\times|S||A|$ matrix Π where $\Pi_{(s,s'a)} = \boldsymbol{\pi}_{(sa)}$ if $s' = s$, otherwise 0. That is, Π is a sparse matrix built by placing $|S|$ row blocks of length $|A|$ in a block diagonal fashion, where each row block gives the conditional distribution over actions specified by $\boldsymbol{\pi}$ in a particular state s. Although this representation of Π might appear unnatural, we find it in fact extremely convenient in our research: from this definition, one can quickly verify that the $|S|\times|S|$ matrix product ΠP gives the *state to state* transition probabilities induced by the policy $\boldsymbol{\pi}$ in the environment P, and the $|S||A|\times|S||A|$ matrix product $P\Pi$ gives the *state-action to state-action* transition probabilities induced by policy $\boldsymbol{\pi}$ in the environment P. We will make repeated use of these two matrix products below.

4. DP with Dual Representations

Dynamic programming methods for solving MDP evaluation and planning problems are typically expressed in terms of the primal value function. We will demonstrate that all of these classical algorithms have natural duals expressed in terms of state and state-action probability distributions. Here the dynamic programming algorithms are organized according to their update types: on-policy update, policy improvement, and off-policy update. In addition both state based and state-action based representations are introduced.

4.1 On-Policy Update

First consider the problem of policy evaluation. Here we assume we are given a fixed policy $\boldsymbol{\pi}$, and wish to compute either its value function in the primal or its discounted visitation distribution in the dual.

4.1.1 State Based Policy Evaluation

First let us examine the standard state based policy evaluation.

Primal Representation. In the primal view, the role of policy evaluation is to recover the *value function*, which is defined to be the expected sum of future discounted rewards. We can now express this definition in vector-matrix form

$$\mathbf{v} \;=\; \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi \mathbf{r} \tag{5}$$

As is well known and easy to verify, this infinite series satisfies a recursive relationship that allows one to recover $\mathbf{v}$ by solving a linear system of $|S|$ equations on $|S|$ unknowns.

$$\mathbf{v} \;=\; \Pi\mathbf{r} + \sum_{i=1}^{\infty} \gamma^i (\Pi P)^i \Pi \mathbf{r} \;=\; \Pi\mathbf{r} + \gamma(\Pi P) \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi \mathbf{r} \;=\; \Pi\mathbf{r} + \gamma \Pi P \mathbf{v} \tag{6}$$

For a given policy Π, the on-policy operator $\mathcal{O}$ is defined as

$$\mathcal{O}\mathbf{v} \;=\; \Pi(\mathbf{r} + \gamma P \mathbf{v})$$

It brings the current representation closer to satisfying the policy-specific Bellman equation (see Equation 6). Therefore, we can have a solution to Equation 6 by iterating the on-policy operator $\mathcal{O}$. We present an analysis of the convergence properties of the on-policy operator in Section 4.4.1 below.

Dual Representation. In the dual form of policy evaluation, one needs to recover a probability distribution over states that has a meaningful correspondence to the long run discounted reward achieved by the policy. Such a correspondence can be achieved by recovering the following probability distribution over states implicitly defined as

$$\mathbf{c}^\top \;=\; (1-\gamma)\boldsymbol{\mu}^\top \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \tag{7}$$

This infinite series satisfies a recursive relationship that allows one to recover $\mathbf{c}$ by solving a linear system of $|S|$ equations on $|S|$ unknowns.

$$\mathbf{c}^\top \;=\; (1-\gamma)\boldsymbol{\mu}^\top + (1-\gamma)\boldsymbol{\mu}^\top \sum_{i=1}^{\infty} \gamma^i (\Pi P)^i \;=\; (1-\gamma)\boldsymbol{\mu}^\top + \gamma \mathbf{c}^\top \Pi P \tag{8}$$

It can be easily verified that Equation 7 defines a probability distribution over states. In addition, the distribution $\mathbf{c}$ also allows one to easily compute the expected discounted return of the policy $\boldsymbol{\pi}$ (see Lemma 3).

Lemma 2 $\mathbf{c}^\top \mathbf{1} = 1$

Proof We know that the definition of $\mathbf{c}$ satisfies Equation 8, then we have $\mathbf{c}^\top \mathbf{1} = (1-\gamma)\boldsymbol{\mu}^\top \mathbf{1} + \gamma \mathbf{c}^\top \Pi P \mathbf{1}$. Since ΠP is row normalized and $\boldsymbol{\mu}$ is a probability distribution, $\mathbf{c}^\top \mathbf{1} = (1-\gamma)1 + \gamma \mathbf{c}^\top \mathbf{1}$ holds. Rearranging the above equation and since $0 \leq \gamma < 1$, we obtain $\mathbf{c}^\top \mathbf{1} = 1$. ■

Lemma 3 $(1-\gamma)\boldsymbol{\mu}^\top \mathbf{v} = \mathbf{c}^\top \Pi \mathbf{r}$

Proof Plugging the definition of $\mathbf{v}$ (see Equation 5) into the left of the above lemma yields $(1-\gamma)\boldsymbol{\mu}^\top \mathbf{v} = (1-\gamma)\boldsymbol{\mu}^\top \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi \mathbf{r}$ Plugging the definition of $\mathbf{c}$ (see Equation 7) into the right of the above lemma yields $\mathbf{c}^\top \Pi \mathbf{r} = (1-\gamma)\boldsymbol{\mu}^\top \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi \mathbf{r}$. Therefore, $(1-\gamma)\boldsymbol{\mu}^\top \mathbf{v} = \mathbf{c}^\top \Pi \mathbf{r}$. ■

Thus, a dual form of policy evaluation can be conducted by recovering $\mathbf{c}$ from Equation 8. The expected discounted reward obtained by policy $\boldsymbol{\pi}$ starting in the initial state distribution $\boldsymbol{\mu}$ can then be computed by $\mathbf{c}^\top \Pi \mathbf{r}/(1-\gamma)$ according to Lemma 3. In principle, this gives a valid form of policy evaluation in a dual representation. However, below we will find that merely recovering the state distribution $\mathbf{c}$ is inadequate for *policy improvement* (see Section 4.2), since there is no apparent way to improve $\boldsymbol{\pi}$ given access to $\mathbf{c}$. Thus, we are compelled to extend the dual representation to a richer representation that avoids an implicit dependence on the initial distribution $\boldsymbol{\mu}$.

Consider the following definition for an $|S| \times |S|$ matrix

$$M \;=\; (1-\gamma) \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \tag{9}$$

This infinite series satisfies a recursive relationship that allows one to recover M by solving a linear system of $|S|$ equations on $|S|$ unknowns.

$$M \;=\; (1-\gamma) I + (1-\gamma) \sum_{i=1}^{\infty} \gamma^i (\Pi P)^i \;=\; (1-\gamma) I + \gamma \Pi P M \tag{10}$$

Based on the above recursive relationship, each row of M can be written as

$$M = \begin{pmatrix} {m_1}^\top \\ {m_2}^\top \\ {m_3}^\top \\ \vdots \\ {m_{|S|}}^\top \end{pmatrix} = \begin{pmatrix} (1-\gamma){\mathbf{e}_1}^\top + \gamma \Pi P {m_1}^\top \\ (1-\gamma){\mathbf{e}_2}^\top + \gamma \Pi P {m_2}^\top \\ (1-\gamma){\mathbf{e}_3}^\top + \gamma \Pi P {m_3}^\top \\ \vdots \\ (1-\gamma){\mathbf{e}_{|S|}}^\top + \gamma \Pi P {m_{|S|}}^\top \end{pmatrix}$$

where $\mathbf{e}_s$ $(s = 1, \ldots, |S|)$ is a vector of all zeros except for a 1 in the s^{th} position. The matrix M that satisfies this linear relationship is similar to $\mathbf{c}^\top$, in that each row is a probability distribution (Lemma 4 below) and the entries $M_{(s,s')}$ correspond to the probability of discounted state visits to s' for a policy $\boldsymbol{\pi}$ starting in state s. However, unlike $\mathbf{c}^\top$, M drops the dependence on $\boldsymbol{\mu}$ and obtains a close relationship with $\mathbf{v}$ (Theorem 6 below).

Lemma 4 $M\mathbf{1} = \mathbf{1}$

Proof By the fact that $(\Pi P)^i \mathbf{1} = \mathbf{1}$ since ΠP is row normalized, we have $M\mathbf{1} = (1-\gamma)\sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \mathbf{1} = (1-\gamma)\sum_{i=0}^{\infty} \gamma^i \mathbf{1} \;=\; \mathbf{1}$. ■

Lemma 5 $\mathbf{c}^\top = \boldsymbol{\mu}^\top M$

Proof From the definition of $\mathbf{c}$, (see Equation 7), $\mathbf{c}^\top = (1-\gamma)\boldsymbol{\mu}^\top \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i$. Plugging the definition of M (see Equation 9) into the right of the above lemma yields $\boldsymbol{\mu}^\top M = \boldsymbol{\mu}^\top (1-\gamma) \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i$ Thus, $\mathbf{c}^\top = \boldsymbol{\mu}^\top M$. ■

Interestingly, Lemmas 4 and 5 show that M is a variant of Dayan's "successor representation" proposed in (Dayan, 1993), but here extended to the infinite horizon discounted case. Moreover, not only is M a matrix of probability distributions over states, it allows one to easily recover the state values of the policy $\boldsymbol{\pi}$.

Similarly, the on-policy operator for a given policy Π is

$$\mathcal{O}M \quad = \quad (1-\gamma)I + \gamma M \Pi P \tag{11}$$

It brings the current representation closer to satisfying the policy-specific Bellman equation (see Equation 10). Therefore, we can have a solution to Equation 10 by iterating the on-policy operator $\mathcal{O}$ (see Section 4.4.1 for convergence results).

Relationship between Primal and Dual Representations. There exists a further connection between state value function $\mathbf{v}$ and state visit distribution M.

Theorem 6 $(1-\gamma)\mathbf{v} = M\Pi\mathbf{r}$

Proof Plugging the definition of $\mathbf{v}$ (see Equation 5) into the left of the above theorem yields $(1-\gamma)\mathbf{v} = (1-\gamma)\sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi \mathbf{r}$. Plugging the definition of M (see Equation 9) into the right of the above theorem yields $M\Pi\mathbf{r} = \left[(1-\gamma)\sum_{i=0}^{\infty} \gamma^i (\Pi P)^i\right] \Pi\mathbf{r}$. Thus, $(1-\gamma)\mathbf{v} = M\Pi\mathbf{r}$. ■

A dual form of policy evaluation can be conducted by recovering M from Equation 10. Then at any time, an equivalent representation to $\mathbf{v}$ can be recovered by $M\Pi\mathbf{r}/(1-\gamma)$, as shown in the above theorem.

4.1.2 State-Action Based Policy Evaluation

Although state based policy evaluation methods like those outlined above are adequate for assessing a given policy, and eventually for formulating DP algorithms, for the RL algorithms below we will generally need to maintain joint *state-action* based evaluations.

Primal Representation. In the primal representation, the state-action value function can be expressed as an $|S||A| \times 1$ vector

$$\mathbf{q} \quad = \quad \sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \mathbf{r} \tag{12}$$

This state-action value function is closely related to the previous state value function (see Equation 6) and satisfies a similar recursive relation.

$$\mathbf{q} \quad = \quad \mathbf{r} + \gamma P \Pi \mathbf{q} \tag{13}$$

For a given policy Π, the on-policy operator is defined as

$$\mathcal{O}\mathbf{q} \quad = \quad \mathbf{r} + \gamma P \Pi \mathbf{q} \tag{14}$$

Iterating the on-policy operator $\mathcal{O}$ gives a solution to Equation 13, which we show formally in Section 4.4.1.

Dual Representation. To develop a dual form of state-action policy evaluation, we use the dual LP representation introduced in Section 3, and represent a probability distribution over state-action pairs that has a useful correspondence to the long run expected discounted rewards achieved by the policy

$$\mathbf{d}^\top \quad = \quad (1-\gamma)\boldsymbol{\nu}^\top \sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \tag{15}$$

where $\boldsymbol{\nu}$ is the initial distribution over the state-action pairs,[1] whose dimension is $|S||A|\times 1$. It can be verified that this defines a probability distribution over state-action pairs. In addtion, the distribution $\mathbf{d}$ also allows one to easily compute the expected discounted return of the policy $\boldsymbol{\pi}$.

This state-action visit distribution is closed related to the previous state visit distribution (see Equation 8) and satisfies a similar recursive relation

$$\mathbf{d}^\top \quad = \quad (1-\gamma)\boldsymbol{\nu}^\top + \gamma \mathbf{d}^\top P \Pi \tag{16}$$

Lemma 7 $\mathbf{d}^\top \mathbf{1} = 1$

Proof Since ΠP is row normalized and $\boldsymbol{\nu}$ is a probability distribution, $\mathbf{d}^\top \mathbf{1} = (1-\gamma)\boldsymbol{\nu}^\top \mathbf{1} + \gamma \mathbf{d}^\top P \Pi \mathbf{1} = (1-\gamma)1 + \gamma \mathbf{d}^\top \mathbf{1}$. Since $0 \leq \gamma < 1$, we have $\mathbf{d}^\top \mathbf{1} = 1$. ■

Lemma 8 $(1-\gamma)\boldsymbol{\nu}^\top \mathbf{q} = \mathbf{d}^\top \mathbf{r}$

Proof It is obvious that $(1-\gamma)\boldsymbol{\nu}^\top \mathbf{q} = \mathbf{d}^\top \mathbf{r}$ by plugging the definitions of $\mathbf{q}$ and $\mathbf{d}$ into the above lemma. ■

A dual form of state-action policy evaluation can be conducted by recovering $\mathbf{d}$ from Equation 16 and computing the expected discounted reward obtained by policy $\boldsymbol{\pi}$ starting in the initial state-action distribution $\boldsymbol{\nu}$ by $\mathbf{d}^\top \mathbf{r}/(1-\gamma)$ (Lemma 8). However, once again we will find that merely recovering the state-action distribution $\mathbf{d}$ is inadequate for *policy improvement* (see Section 4.2), since there is no apparent way to improve $\boldsymbol{\pi}$ given access to $\mathbf{d}$. Thus, again, we extend the dual representation to a richer representation that avoids an implicit dependence on the initial distribution $\boldsymbol{\nu}$.

Consider the following definition for an $|S||A|\times|S||A|$ matrix

$$H \quad = \quad (1-\gamma) \sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \tag{17}$$

1. By the definitions of initial state visit distribution $\boldsymbol{\mu}$ and state-action visit distribution $\boldsymbol{\nu}$, we can have relationship $\boldsymbol{\nu}^\top = \boldsymbol{\mu}^\top \Pi$.

This infinite series satisfies a recursive relationship that allows one to recover H by solving a linear system of $|S||A|$ equations on $|S||A|$ unknowns.

$$H \quad = \quad (1-\gamma)I + \gamma P\Pi H \tag{18}$$

The matrix H that satisfies this linear relation is similar to $\mathbf{d}^\top$, in that each row is a probability distribution and the entries $H_{(sa,s'a')}$ correspond to the probability of discounted state-action visits to $(s'a')$ for a policy $\boldsymbol{\pi}$ starting in state-action pair (sa). However, H drops the dependence on $\boldsymbol{\mu}$ and obtains a close relationship with $\mathbf{q}$ (Theorem 11 below).

Lemma 9 $H\mathbf{1} = \mathbf{1}$

Proof Since $P\Pi$ is row normalized, $(P\Pi)^i\mathbf{1} = \mathbf{1}$. $H\mathbf{1} = (1-\gamma)\sum_{i=0}^{\infty}\gamma^i(P\Pi)^i\mathbf{1} = (1-\gamma)\sum_{i=0}^{\infty}\gamma^i\mathbf{1} = \mathbf{1}$. ■

Lemma 10 $\mathbf{d}^\top = \boldsymbol{\nu}^\top H$

Proof It is obvious that $\mathbf{d}^\top = \boldsymbol{\nu}^\top H$ by plugging the definitions of $\mathbf{d}$ and H into the above lemma. ■

For a given policy Π, the on-policy operator $\mathcal{O}$ is defined as

$$\mathcal{O}H \quad = \quad (1-\gamma)I + \gamma P\Pi H \tag{19}$$

It brings the current representation closer to satisfying the policy-specific Bellman equation (see Equation 18). Iterating the on-policy operator $\mathcal{O}$ gives a solution to Equation 18, which we show formally in Section 4.4.1.

Relationship between Primal and Dual Representations. There is a strong relationship between state-action value function $\mathbf{q}$ and state-action visit distribution H as follows.

Theorem 11 $(1-\gamma)\mathbf{q} = H\mathbf{r}$

Proof Plugging the definition of $\mathbf{q}$ into the left of the above theorem yields

$$(1-\gamma)\mathbf{q} \quad = \quad (1-\gamma)\sum_{i=0}^{\infty}\gamma^i(P\Pi)^i\mathbf{r}$$

Plugging the definition of H into the right of the above theorem yields

$$H\mathbf{r} \quad = \quad \left[(1-\gamma)\sum_{i=0}^{\infty}\gamma^i(P\Pi)^i\right]\mathbf{r}$$

Thus,

$$(1-\gamma)\mathbf{q} \quad = \quad H\mathbf{r}$$

■

A dual form of state-action policy evaluation can be conducted by recovering H from its Bellman equation (see Equation 18). Then at any time, an equivalent representation to $\mathbf{q}$ can be recovered by $H\mathbf{r}/(1-\gamma)$. Note there is a many to one relationship between the dual and primal representations because the number of variables in H ($|S||A| \times |S||A|$) is more than the number of the constraints given by their relation, as shown in Theorem 11.

Relationship between State and State-action Based Representations. Finally, one can relate the state and state-action matrix representations defined above to each other. We can have the relationship between state value function $\mathbf{v}$ and state-action value function $\mathbf{q}$ in the primal representation as follows.

Lemma 12 $\mathbf{v} = \Pi\mathbf{q}$

Proof By the definitions of $\mathbf{v}$ and $\mathbf{q}$, $\Pi\mathbf{q} = \Pi \sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \mathbf{r} = \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi\mathbf{r} = \mathbf{v}$. ■

The relationship between state visit distribution M and state-action visit distributions H in the dual representation is as follows.

Lemma 13 $M\Pi = \Pi H$

Proof By the definitions of M and H, we have $M\Pi = (1-\gamma)\sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi = (1-\gamma)\sum_{i=0}^{\infty} \gamma^i \Pi (\Pi P)^i = \Pi(1-\gamma)\sum_{i=0}^{\infty} \gamma^i (\Pi P)^i = \Pi H$. ■

Thus, to this point, we have developed new dual representations that can form the basis for state based and state-action based policy evaluation, respectively. These are defined in terms of state distributions and state-action distributions, and do not require value functions to be computed. In the next section, we will examine the convergence properties of on-policy updates with the dual representations.

4.2 Policy Improvement

The next step is to consider mechanisms for policy improvement, which combined with policy evaluation form policy iteration algorithms capable of solving MDP planning problems.

4.2.1 Primal Representation

The standard primal policy improvement update is well known. Given a current policy $\boldsymbol{\pi}$, whose state value function $\mathbf{v}$ or state-action value function $\mathbf{q}$ have already been determined, one can derive an improved policy $\boldsymbol{\pi}'$ via the update

$$a^*(s) = \arg\max_a \mathbf{q}_{(sa)} \quad (20)$$

$$= \arg\max_a \mathbf{r}_{(sa)} + \gamma P_{(sa,:)}\mathbf{v} \quad (21)$$

The imporved policy $\boldsymbol{\pi}'_{(sa)} = 1$ if $a = a^*(s)$, otherwise 0. The subsequent "policy improvement theorem" verifies that this update leads to an improved policy.

Theorem 14 $\Pi\mathbf{q} \leq \Pi'\mathbf{q}$ *implies* $\mathbf{v} \leq \mathbf{v}'$

Proof Since $\mathbf{v} = \Pi\mathbf{q}$ (see Lemma 12) and $\mathbf{q} = \mathbf{r} + \gamma P\Pi\mathbf{q}$ (see Equation 13)

$$\mathbf{q} \;=\; \mathbf{r} + \gamma P\mathbf{v}$$

Plugging the above relation into the assumption $\Pi\mathbf{q} \leq \Pi'\mathbf{q}$, we have

$$\Pi(\mathbf{r} + \gamma P\mathbf{v}) \;\leq\; \Pi'(\mathbf{r} + \gamma P\mathbf{v}) \implies \Pi\mathbf{r} + \gamma\Pi P\mathbf{v} \;\leq\; \Pi'\mathbf{r} + \gamma\Pi' P\mathbf{v}$$

Using $\mathbf{v} = \Pi\mathbf{r} + \gamma\Pi P\mathbf{v}$ (see Equation 6) recursively, the above inequality relation becomes

$$\begin{aligned}
\mathbf{v} &\leq \Pi'\mathbf{r} + \gamma\Pi' P\mathbf{v} = \Pi'\mathbf{r} + \gamma\Pi' P\Pi(\mathbf{r} + \gamma P\mathbf{v}) \\
&\leq \Pi'\mathbf{r} + \gamma\Pi' P\Pi'(\mathbf{r} + \gamma P\mathbf{v}) = \Pi'\mathbf{r} + \gamma\Pi' P\Pi'\mathbf{r} + \gamma^2(\Pi' P)^2\mathbf{v} \\
&\vdots \\
&= \sum_{i=0}^{\infty} \gamma^i (\Pi' P)^i \Pi'\mathbf{r}
\end{aligned}$$

From the definition of $\mathbf{v}$, $\sum_{i=0}^{\infty} \gamma^i (\Pi' P)^i \Pi'\mathbf{r} = \mathbf{v}'$ holds. Therefore, $\mathbf{v} \leq \mathbf{v}'$. ■

4.2.2 Dual Representation

The above development can be parallelled in the dual by first defining an analogous policy update and proving an analogous policy improvement theorem. Given a current policy $\boldsymbol{\pi}$, in the dual one can derive an improved policy $\boldsymbol{\pi}'$ by the update

$$\begin{aligned}
a^*(s) &= \arg\max_a H_{(sa,:)}\mathbf{r} && (22) \\
&= \arg\max_a (1-\gamma)\mathbf{r}_{(sa)} + \gamma P_{(sa,:)} M\Pi\mathbf{r} && (23) \\
\boldsymbol{\pi}'_{(sa)} &= \begin{cases} 1 & \text{if } a = a^*(s) \\ 0 & \text{if } a \neq a^*(s) \end{cases}
\end{aligned}$$

In fact, by Theorem 11, the two policy updates given in Equation 20 and Equation 22 respectively, must lead to the same resulting policy $\boldsymbol{\pi}'$. Therefore, not surprisingly, we have an analogous policy improvement theorem in this case.

Theorem 15 *$\Pi H\mathbf{r} \leq \Pi' H\mathbf{r}$ implies $M\Pi\mathbf{r} \leq M'\Pi'\mathbf{r}$*

Proof Plugging $M\Pi = \Pi H$ (see Lemma 13) and $H = (1-\gamma)I + \gamma P\Pi H$ (see Equation 18) into the assumption $\Pi H\mathbf{r} \leq \Pi' H\mathbf{r}$, we have

$$M\Pi\mathbf{r} \;\leq\; (1-\gamma)\Pi'\mathbf{r} + \gamma\Pi' P\Pi H\mathbf{r}$$

Using the assumption $\Pi H\mathbf{r} \leq \Pi' H\mathbf{r}$ and Equation 18 recursively in the above relation,

$$\begin{aligned}
(1-\gamma)\Pi'\mathbf{r} + \gamma\Pi' P\Pi H\mathbf{r} &\leq (1-\gamma)\Pi'\mathbf{r} + \gamma\Pi' P\Pi' H\mathbf{r} \\
&= (1-\gamma)\Pi'\mathbf{r} + \gamma\Pi' P\Pi' \left[(1-\gamma)I + \gamma P\Pi H\right]\mathbf{r} \\
&= (1-\gamma)\Pi'\mathbf{r} + (1-\gamma)\gamma\Pi' P\Pi'\mathbf{r} + \gamma^2(\Pi' P)^2\Pi H\mathbf{r} \\
&\leq (1-\gamma)\Pi'\mathbf{r} + (1-\gamma)\gamma\Pi' P\Pi'\mathbf{r} + \gamma^2(\Pi' P)^2\Pi' H\mathbf{r} \\
&\vdots \\
&= (1-\gamma)\sum_{i=0}^{\infty}\gamma^i(\Pi' P)^i\Pi'\mathbf{r}
\end{aligned}$$

From the definition of M (see Equation 9), we have $(1-\gamma)\sum_{i=0}^{\infty}\gamma^i(\Pi' P)^i\Pi'\mathbf{r} = M'\Pi'\mathbf{r}$. Therefore, $M\Pi\mathbf{r} \leq M'\Pi'\mathbf{r}$. ■

Thus, a dual state policy iteration algorithm can be completely expressed in terms of the dual representation M, incorporating both dual policy evaluation (see Equation 10) and dual policy improvement (see Equation 23) leading to an equivalent result to the standard primal policy iteration algorithm based on $\mathbf{v}$ (see Equation 6) and primal policy improvement (see Equation 21).

4.3 Off-Policy Update

The off-policy update is prominent in dynamic programming (e.g., value iteration) and reinforcement learning algorithms (e.g., Q-learning). The off-policy update $\mathcal{M}$ is different from on-policy update $\mathcal{O}$ in that it is neither linear nor defined by any reference policy, but instead applies a *greedy max* update to the current estimates.

4.3.1 Primal Representations

In the primal case, Bellman iteration of state value function corresponds to the well known value iteration algorithm, which is based on the off-policy update $\mathcal{M}$ for state value function

$$\mathcal{M}\mathbf{v} \;=\; \Pi^*[\mathbf{r} + \gamma P\mathbf{v}] \quad \text{where } \Pi^*[\mathbf{r} + \gamma P\mathbf{v}]_{(s)} \;=\; \max_a \mathbf{r}_{(sa)} + \gamma P_{(sa,:)}\mathbf{v} \tag{24}$$

The goal of this greedy update is to bring the representation $\mathbf{v}$ closer to satisfying the optimal-policy Bellman equation $\mathbf{v} = \Pi^*[\mathbf{r} + \gamma P\mathbf{v}]$.

Similarly, the off-policy update operator $\mathcal{M}$ for state-action value function $\mathbf{q}$ is

$$\mathcal{M}\mathbf{q} \;=\; \mathbf{r} + \gamma P\Pi^*[\mathbf{q}] \quad \text{where } \Pi^*[\mathbf{q}]_{(s)} \;=\; \max_a \mathbf{q}_{(sa)} \tag{25}$$

The goal of this greedy update is to bring the representation $\mathbf{q}$ closer to satisfying the optimal-policy Bellman equation $\mathbf{q} = \mathbf{r} + \gamma P\Pi^*[\mathbf{q}]$.

4.3.2 Dual Representations

In the dual, Bellman iteration bypass the explicit representation of the value of a policy, and attempt to update the evaluation of the optimal policy implicitly. An analogous off-policy

update $\mathcal{M}$ for the state visit distribution can be defined as

$$\begin{aligned} \mathcal{M}M &= (1-\gamma)I + \gamma\Pi^*_{\mathbf{r}}[PM], \quad \text{where} \\ \Pi^*_{\mathbf{r}}[PM]_{(s)} &= \max_a [PM\Pi\mathbf{r}]_{(sa)} = \max_a \sum_{(s'a')} P_{(sa,s'a')} M_{(ss')} \Pi_{(s,s'a')} \mathbf{r}_{(s'a')} \end{aligned} \tag{26}$$

The goal of this greedy update is to bring the representation M closer to satisfying the optimal-policy Bellman equation $M = (1-\gamma)I + \gamma\Pi^*_{\mathbf{r}}[PM]$.

Similarly, the max-policy update operator $\mathcal{M}$ for state-action visit distribution H is defined as

$$\begin{aligned} \mathcal{M}H &= (1-\gamma)I + \gamma P\Pi^*_{\mathbf{r}}[H], \quad \text{where} \\ \Pi^*_{\mathbf{r}}[H]_{(s)} &= \max_a [H\mathbf{r}]_{(sa)} = \max_a \sum_{(s'a')} H_{(sa,s'a')} \mathbf{r}_{(s'a')} \end{aligned} \tag{27}$$

The goal of this greedy update is to bring the representation H closer to satisfying the optimal-policy Bellman equation $H = (1-\gamma)I + \gamma P\Pi^*_{\mathbf{r}}[H]$.

Note that a dual form of Bellman iteration algorithms need not refer to the primal value functions at all. Nevertheless, the off-policy update of $\mathbf{q}$ (see Equation 25) and the off-policy update of H (see Equation 27) behave equivalently because of their relation $(1-\gamma)\mathbf{q} = H\mathbf{r}$ (see Theorem 11 in Section 4.1.2).

4.4 Convergence Analysis

We first investigate whether dynamic programming operators with the dual representations exhibit the same convergence properties to their primal counterparts. These questions will be answered in the affirmative way by proof. In the tabular case, dynamic programming algorithms can be expressed by operators that are successively applied to current approximations (vectors in the primal case, matrices in the dual), to bring them closer to a target solution; namely, the fixed point of a desired Bellman equation. Consider two standard operators, the on-policy update and the max-policy update. Recall that for a given policy Π, the on-policy operator $\mathcal{O}$ is defined as

$$\mathcal{O}\mathbf{q} = \mathbf{r} + \gamma P\Pi\mathbf{q} \qquad \text{and} \qquad \mathcal{O}H = (1-\gamma)I + \gamma P\Pi H,$$

for the primal and dual cases respectively. The goal of the on-policy update is to bring current representations closer to satisfying the policy-specific Bellman equations, $\mathbf{q} = \mathbf{r} + \gamma P\Pi\mathbf{q}$ and $H = (1-\gamma)I + \gamma P\Pi H$.

The max-policy operator $\mathcal{M}$ is different in that it is neither linear nor defined by any reference policy, but instead applies a greedy max update to the current approximations

$$\mathcal{M}\mathbf{q} = \mathbf{r} + \gamma P\Pi^*[\mathbf{q}] \qquad \text{and} \qquad \mathcal{M}H = (1-\gamma)I + \gamma P\Pi^*_r[H],$$

where $\Pi^*[\mathbf{q}]_{(s)} = \max_a \mathbf{q}_{(sa)}$ and $\Pi^*_r[H]_{(s)} = \max_a [H\mathbf{r}]_{(sa)} = \max_a \sum_{s'a'} H_{(sa,s'a')} \mathbf{r}_{(s'a')}$. The goal of this greedy update is to bring the representations closer to satisfying the optimal-policy Bellman equations $\mathbf{q} = \mathbf{r} + \gamma P\Pi^*[\mathbf{q}]$ and $H = (1-\gamma)I + \gamma P\Pi^*_r[H]$.

4.4.1 On-policy convergence

For the on-policy operator $\mathcal{O}$, convergence to the Bellman fixed point is easily proved in the primal case, by establishing a contraction property of $\mathcal{O}$ with respect to a specific norm on $\mathbf{q}$ vectors. One defines a weighted 2-norm with weights given by the stationary distribution determined by the policy Π with respect to the transition model P. Let $\mathbf{z} \geq 0$ be a vector such that $\mathbf{z}^\top P\Pi = \mathbf{z}^\top$; that is, $\mathbf{z}$ is the stationary state-action visit distribution for $P\Pi$. The norm is defined as $\|\mathbf{q}\|_{\mathbf{z}}{}^2 = \mathbf{q}^\top Z\mathbf{q} = \sum_{(sa)} \mathbf{z}_{(sa)}\mathbf{q}^2_{(sa)}$. where $Z = \mathrm{diag}(\mathbf{z})$. It can be shown that $\|P\Pi\mathbf{q}\|_{\mathbf{z}} \leq \|\mathbf{q}\|_{\mathbf{z}}$ and $\|\mathcal{O}\mathbf{q}_1 - \mathcal{O}\mathbf{q}_2\|_{\mathbf{z}} \leq \gamma\|\mathbf{q}_1 - \mathbf{q}_2\|_{\mathbf{z}}$ (see (Tsitsiklis and Van Roy, 1997)). Crucially, for this norm, a state-action transition is not an expansion (Tsitsiklis and Van Roy, 1997). By the contraction map fixed point theorem (Bertsekas, 1995) there exists a unique fixed point of $\mathcal{O}$ in the space of vectors $\mathbf{q}$. Therefore, repeated applications of the on-policy operator converge to a vector $\mathbf{q}_\Pi$ such that $\mathbf{q}_\Pi = \mathcal{O}\mathbf{q}_\Pi$; that is, $\mathbf{q}_\Pi$ satisfies the policy based Bellman equation.

Analogously, for the dual representation H, one can establish convergence of the on-policy operator by first defining an approximate weighted norm over matrices and then verifying that $\mathcal{O}$ is a contraction with respect to this norm. Define

$$\|H\|_{\mathbf{z},\mathbf{r}}{}^2 \;=\; \|H\mathbf{r}\|_{\mathbf{z}}{}^2 \;=\; \sum_{(sa)} \mathbf{z}_{(sa)} \Big(\sum_{(s'a')} H_{(sa,s'a')}\mathbf{r}_{(s'a')} \Big)^2 \tag{28}$$

It is easily verified that this definition satisfies the property of a pseudo-norm, and in particular, satisfies the triangle inequality. This weighted 2-norm is defined with respect to the stationary distribution $\mathbf{z}$, but also the reward vector $\mathbf{r}$. Thus, the magnitude of a row normalized matrix is determined by the magnitude of the weighted reward expectations it induces.

Interestingly, this definition allows us to establish the same non-expansion and contraction results as the primal case. For example, state-action transitions remain a non-expansion.

Lemma 16 $\|P\Pi H\|_{\mathbf{z},\mathbf{r}} \leq \|H\|_{\mathbf{z},\mathbf{r}}$

Proof From Jensen's inequality, we have

$$\begin{aligned}
\|P\Pi(H\mathbf{r})\|_{\mathbf{z}}{}^2 &= \sum_{(sa)} \mathbf{z}_{(sa)} \left(\sum_{(s'a')} [P\Pi]_{(sa,s'a')} (H\mathbf{r})_{(s'a')} \right)^2 \leq \sum_{(sa)} \mathbf{z}_{(sa)} \sum_{(s'a')} [P\Pi]_{(sa,s'a')} (H\mathbf{r})^2_{(s'a')} \\
&= \sum_{(s'a')} (H\mathbf{r})^2_{(s'a')} \sum_{(sa)} [P\Pi]_{(sa,s'a')} \mathbf{z}_{(sa)} = \sum_{(s'a')} (H\mathbf{r})^2_{(s'a')} \mathbf{z}_{(s'a')} = \|H\mathbf{r}\|_{\mathbf{z}}{}^2
\end{aligned}$$

Together with the definition (Equation 28), $\|P\Pi H\|_{\mathbf{z},\mathbf{r}} = \|P\Pi(H\mathbf{r})\|_{\mathbf{z}} \leq \|H\mathbf{r}\|_{\mathbf{z}} = \|H\|_{\mathbf{z},\mathbf{r}}$. ■

Moreover, the on-policy operator is a contraction with respect to $\|\cdot\|_{\mathbf{z},\mathbf{r}}$.

Lemma 17 $\|\mathcal{O}H_1 - \mathcal{O}H_2\|_{\mathbf{z},\mathbf{r}} \leq \gamma\|H_1 - H_2\|_{\mathbf{z},\mathbf{r}}$

Proof By the definition of on-policy operator $\mathcal{O}$, we have

$$\begin{aligned}\|\mathcal{O}H_1 - \mathcal{O}H_2\|_{\mathbf{z},\mathbf{r}} &= \|(1-\gamma)I + \gamma P\Pi H_1 - (1-\gamma)I - \gamma P\Pi H_2\|_{\mathbf{z},\mathbf{r}} \\ &= \|\gamma P\Pi H_1 - \gamma P\Pi H_2\|_{\mathbf{z},\mathbf{r}} = \gamma\|P\Pi(H_1 - H_2)\|_{\mathbf{z},\mathbf{r}}\end{aligned}$$

Together with Lemma 16, we obtain $\|\mathcal{O}H_1 - \mathcal{O}H_2\|_{\mathbf{z},\mathbf{r}} \leq \gamma\|H_1 - H_2\|_{\mathbf{z},\mathbf{r}}$. ■

Thus, once again by the contraction map fixed point theorem there exists a fixed point of $\mathcal{O}$ among row normalized matrices H, and repeated applications of $\mathcal{O}$ will converge to a matrix H_Π such that $\mathcal{O}H_\Pi = H_\Pi$; that is, H_Π satisfies the policy based Bellman equation for dual representations. This argument shows that on-policy dynamic programming converges in the dual representation, without making direct reference to the primal case. We will use these results below. A simpler argument would have been to reduce the dual to the primal case, which we do now for the max operator.

4.5 Max-policy convergence

The strategy for establishing convergence for the nonlinear max operator is similar to the on-policy case, but involves working with a different norm. Instead of considering a 2-norm weighted by the visit probabilities induced by a fixed policy, one simply uses the max-norm in this case: $\|\mathbf{q}\|_\infty = \max_{(sa)} |q_{(sa)}|$. The contraction property of the $\mathcal{M}$ operator with respect to this norm can then be easily established in the primal case: $\|\mathcal{M}\mathbf{q}_1 - \mathcal{M}\mathbf{q}_2\|_\infty \leq \gamma\|\mathbf{q}_1 - \mathbf{q}_2\|_\infty$ (see (Bertsekas, 1995)). As in the on-policy case, contraction suffices to establish the existence of a unique fixed point of $\mathcal{M}$ among vectors $\mathbf{q}$, and that repeated application of $\mathcal{M}$ converges to this fixed point $\mathbf{q}_*$ such that $\mathcal{M}\mathbf{q}_* = \mathbf{q}_*$.

To establish convergence of the off-policy update in the dual representation, first define the max-norm for state-action visit distribution as

$$\|H\|_\infty = \max_{(sa)} |\sum_{(s'a')} H_{(sa,s'a')}\mathbf{r}_{(s'a')}| \tag{29}$$

Then one can simply reduce the dual to the primal case by appealing to the relationship $(1-\gamma)\mathcal{M}\mathbf{q} = \mathcal{M}H\mathbf{r}$ to prove convergence of $\mathcal{M}H$.

Lemma 18 *If* $(1-\gamma)\mathbf{q} = H\mathbf{r}$*, then* $(1-\gamma)\mathcal{M}\mathbf{q} = \mathcal{M}H\mathbf{r}$.

Proof From the assumption, $(1-\gamma)\mathbf{q} = H\mathbf{r} \implies (1-\gamma)\mathbf{q}_{(sa)} = [H\mathbf{r}]_{(sa)}\ \forall(sa)$. Together with the definitions of Π^* in Equation 25 and 27, we have $\Pi^*[\mathbf{q}] = \Pi^*[H\mathbf{r}] \implies \Pi^*[\mathbf{q}] = \Pi^*_{\mathbf{r}}[H]\mathbf{r}$. Applying $\mathcal{M}$ update on $\mathbf{q}$ (see Equation 25), we get $(1-\gamma)\mathcal{M}\mathbf{q} = (1-\gamma)(\mathbf{r} + \gamma P\Pi^*[\mathbf{q}])$. Substituting $\Pi^*[\mathbf{q}]$ with $\Pi^*_{\mathbf{r}}[H]\mathbf{r}$, we have $(1-\gamma)\mathcal{M}\mathbf{q} = (1-\gamma)(I + \gamma P\Pi^*_{\mathbf{r}}[H])\mathbf{r}$. Note that $\mathcal{M}H = I + \gamma P\Pi^*_{\mathbf{r}}[H]$ (see Equation 27), $(1-\gamma)\mathcal{M}\mathbf{q} = \mathcal{M}H\mathbf{r}$ holds. ■

Thus, given convergence of $\mathcal{M}\mathbf{q}$ to a fixed point $\mathcal{M}\mathbf{q}_* = \mathbf{q}_*$, the same must also hold for $\mathcal{M}H$. However, one subtlety here is that the dual fixed point is not unique. This is not a contradiction because the norm on dual representations $\|\cdot\|_{\mathbf{z},\mathbf{r}}$ is in fact just a pseudo-norm, not a proper norm. That is, the relationship between H and $\mathbf{q}$ is many to one, and several

matrices can correspond to the same $\mathbf{q}$. These matrices form a convex subspace (in fact, a simplex), since if $H_1\mathbf{r} = (1-\gamma)\mathbf{q}$ and $H_2\mathbf{r} = (1-\gamma)\mathbf{q}$ then $(\alpha H_1 + (1-\alpha)H_2)\mathbf{r} = (1-\gamma)\mathbf{q}$ for any α, where furthermore α must be restricted to $0 \leq \alpha \leq 1$ to maintain nonnegativity. The simplex of fixed points $\{H_* : \mathcal{M}H_* = H_*\}$ is given by matrices H_* that satisfy $H_*\mathbf{r} = (1-\gamma)\mathbf{q}_*$.

5. DP with function approximation

Primal and dual updates exhibit strong equivalence in the tabular case, as they should. However, when we begin to consider approximation, differences emerge. We next consider the convergence properties of the dynamic programming operators in the context of linear basis approximation. We focus on the on-policy case here, because, famously, the max operator does not always have a fixed point when combined with approximation in the primal case (de Farias and Van Roy, 2000), and consequently suffers the risk of divergence (Baird, 1995; Sutton and Barto, 1998).

Note that the max operator cannot diverge in the dual case, even with basis approximation, by boundedness alone as we will show through empirical studies in Section 6; although the question of whether max updates always converge in this case remains open. Here we establish that a similar bound on approximation error in the primal case can be proved for the dual approach with respect to the on-policy operator.

In the primal case, linear approximation proceeds by fixing a small set of basis functions, forming a $|S||A|\times k$ matrix Φ, where k is the number of bases. The approximation of $\mathbf{q}$ can be expressed by a linear combination of bases $\hat{\mathbf{q}} = \Phi\mathbf{w}$ where $\mathbf{w}$ is a $k\times 1$ vector of adjustable weights. This is equivalent to maintaining the constraint that $\hat{\mathbf{q}} \in \text{col\_span}(\Phi)$, where col_span is the column span. In the dual, a linear approximation to H can be expressed as $\text{vec}(\hat{H}) = \Psi\mathbf{w}$, where the vec operator creates a column vector from a matrix by stacking the column vectors of the matrix below one another, $\mathbf{w}$ is a $k\times 1$ vector of adjustable weights as it is in the primal case, and Ψ is a $(|S||A|)^2 \times k$ matrix of basis functions. To ensure that $\hat{H}$ remains a nonnegative, row normalized approximation to H, that is, we require $\hat{H} \in \text{simplex}(\Psi) \equiv \{\hat{H} : \text{vec}(\hat{H}) = \Psi\mathbf{w}, \Psi \geq 0, (\mathbf{1}^\top \otimes I)\Psi = \mathbf{1}\mathbf{1}^\top, \mathbf{w} \geq 0, \mathbf{w}^\top\mathbf{1} = \mathbf{1}\}$ where the operator $\otimes$ is the Kronecker product.

In this section, we first introduce operators (a projection operator and a gradient operator) that ensure the approximations stay representable in the given bases. Then we consider their composition with the on-policy update and off-policy update, and analyze their convergence properties. For the composition of the on-policy update and projection operators, we establish a similar bound on approximation error in the dual case as in the primal case.

5.1 Projection Operator

Recall that in the primal, the action value function $\mathbf{q}$ is approximated by a linear combination of bases in Φ. Unfortunately, there is no reason to expect $\mathcal{O}\mathbf{q}$ or $\mathcal{M}\mathbf{q}$ to stay in the column span of Φ, so a best approximation is required. The subtlety resolved by Tsitsiklis and Van Roy (1997) is to identify a particular form of best approximation—weighted least squares—that ensures convergence is still achieved when combined with the on-policy operator $\mathcal{O}$. Unfortunately, the fixed point of this combined update operator is not guar-

anteed to be the best representable approximation of $\mathcal{O}$'s fixed point, $\mathbf{q}_\Pi$. Nevertheless, a bound can be proven on how close this altered fixed point is to the best representable approximation.

We summarize a few details that will be useful below: First, the best least squares approximation is computed with respect to the distribution $\mathbf{z}$. The map from a general $\mathbf{q}$ vector onto its best approximation in col_span(Φ) is defined by another operator, $\mathcal{P}$, which projects $\mathbf{q}$ into the column span of Φ, $\hat{\mathbf{q}} = \mathcal{P}\mathbf{q} = \Phi(\Phi^\top Z\Phi)^{-1}\Phi^\top Z\mathbf{q}$ where $\hat{\mathbf{q}}$ is an approximation for value function $\mathbf{q}$ and recall that $Z = \text{diag}(\mathbf{z})$. The important property of this weighted projection is that it is a non-expansion operator in $\|\cdot\|_\mathbf{z}$, i.e., $\|\mathcal{P}\mathbf{q}\|_\mathbf{z} \leq \|\mathbf{q}\|_\mathbf{z}$, which can be easily obtained from the generalized Phythagorean theorem. Approximate dynamic programming then proceeds by composing the two operators—the on-policy update $\mathcal{O}$ with the subspace projection $\mathcal{P}$ — essentially computing the best representable approximation of the one step update. This combined operator is guaranteed to converge, since composing a non-expansion with a contraction is still a contraction, that is, $\|\mathbf{q}_+ - \mathbf{q}_\Pi\|_\mathbf{z} \leq \frac{1}{1-\gamma}\|\mathbf{q}_\Pi - \mathcal{P}\mathbf{q}_\Pi\|_\mathbf{z}$ (Tsitsiklis and Van Roy, 1997).

Linear function approximation in the dual case is a bit more complicated because matrices are being represented, not vectors, and moreover the matrices need to satisfy row normalization and nonnegativity constraints. Nevertheless, a very similar approach to the primal case can be successfully applied. Recall that in the dual, the state-action visit distribution H is approximated by a linear combination of bases in Ψ. As in the primal case, there is no reason to expect that an update like $\mathcal{O}H$ should keep the matrix in the simplex. Therefore, a projection operator must be constructed that determines the best representable approximation to $\mathcal{O}H$. One needs to be careful to define this projection with respect to the right norm to ensure convergence. Here, the pseudo-norm $\|\cdot\|_{\mathbf{z},\mathbf{r}}$ defined in Equation 28 suits this purpose. Define the weighted projection operator $\mathcal{P}$ over matrices

$$\mathcal{P}H \quad = \quad \underset{\hat{H}\in\text{simplex}(\Psi)}{\text{argmin}} \quad \|H - \hat{H}\|_{\mathbf{z},\mathbf{r}}{}^2 \tag{30}$$

The projection could be obtained by solving the above quadratic program. A key result is that this projection operator is a non-expansion with respect to the pseudo-norm $\|\cdot\|_{\mathbf{z},\mathbf{r}}$.

Theorem 19 $\|\mathcal{P}H\|_{\mathbf{z},\mathbf{r}} \leq \|H\|_{\mathbf{z},\mathbf{r}}$

Proof The easiest way to prove the theorem is to observe that the projection operator $\mathcal{P}$ is really a composition of three orthogonal projections: first, onto the linear subspace span(Ψ), then onto the subspace of row normalized matrices span(Ψ) $\cap \{H : H\mathbf{1} = \mathbf{1}\}$, and finally onto the space of nonnegative matrices span(Ψ)$\cap\{H : H\mathbf{1} = \mathbf{1}\}\cap\{H : H \geq 0\}$. Note that the last projection into the nonnegative halfspace is equivalent to a projection into a linear subspace for some hyperplane tangent to the simplex. Each one of these projections is a non-expansion in $\|\cdot\|_{\mathbf{z},\mathbf{r}}$ in the same way: a generalized Pythagorean theorem holds. Consider just one of these linear projections $\mathcal{P}_1$

$$\begin{aligned}\|H\|_{\mathbf{z},\mathbf{r}}{}^2 &= \|\mathcal{P}_1 H + H - \mathcal{P}_1 H\|_{\mathbf{z},\mathbf{r}}{}^2 = \|\mathcal{P}_1 H\mathbf{r} + H\mathbf{r} - \mathcal{P}_1 H\mathbf{r}\|_\mathbf{z}{}^2 \\ &= \|\mathcal{P}_1 H\mathbf{r}\|_\mathbf{z}{}^2 + \|H\mathbf{r} - \mathcal{P}_1 H\mathbf{r}\|_\mathbf{z}{}^2 = \|\mathcal{P}_1 H\|_{\mathbf{z},\mathbf{r}}{}^2 + \|H - \mathcal{P}_1\|_{\mathbf{z},\mathbf{r}}{}^2\end{aligned}$$

Since the overall projection is just a composition of non-expansions, it must be a non-expansion. ■

As in the primal, approximate dynamic programming can be implemented by composing the on-policy update $\mathcal{O}$ with the projection operator $\mathcal{P}$. Since $\mathcal{O}$ is a contraction and $\mathcal{P}$ a non-expansion, $\mathcal{PO}$ must also be a contraction, and it then follows that it has a fixed point. Note that, as in the tabular case, this fixed point is only unique up to $H\mathbf{r}$-equivalence, since the pseudo-norm $\|\cdot\|_{\mathbf{z},\mathbf{r}}$ does not distinguish H_1 and H_2 such that $H_1\mathbf{r} = H_2\mathbf{r}$. Here too, the fixed point is actually a simplex of equivalent solutions. For simplicity, we denote the simplex of fixed points for $\mathcal{PO}$ by some representative $H_+ = \mathcal{PO}H_+$. Finally, we can recover an approximation bound that is analogous to the primal bound, which bounds the approximation error between H_+ and the best representable approximation to the on-policy fixed point $H_\Pi = \mathcal{O}H_\Pi$.

Theorem 20 $\|H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}} \leq \frac{1}{1-\gamma}\|\mathcal{P}H_\Pi - H_\Pi\|_{\mathbf{z},\mathbf{r}}$

Proof First note that $\|H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}} = \|H_+ - \mathcal{P}H_\Pi + \mathcal{P}H_\Pi - H_\Pi\|_{\mathbf{z},\mathbf{r}} \leq \|H_+ - \mathcal{P}H_\Pi\|_{\mathbf{z},\mathbf{r}} + \|\mathcal{P}H_\Pi - H_\Pi\|_{\mathbf{z},\mathbf{r}}$ by generalized Phythagorean theorem. Notice that $H_+ = \mathcal{PO}H_+$ and $\mathcal{P}$ is a non-expansion operator, we have $\|H_+ - \mathcal{P}H_\Pi\|_{\mathbf{z},\mathbf{r}} = \|\mathcal{PO}H_+ - \mathcal{P}H_\Pi\|_{\mathbf{z},\mathbf{r}} \leq \|\mathcal{O}H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}}$. Since $H_\Pi = \mathcal{O}H_\Pi$ and by Lemma 17, $\|\mathcal{O}H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}} = \|\mathcal{O}H_+ - \mathcal{O}H_\Pi\|_{\mathbf{z},\mathbf{r}} \leq \gamma\|H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}}$. Thus, $(1-\gamma)\|H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}} \leq \|\mathcal{P}H_\Pi - H_\Pi\|_{\mathbf{z},\mathbf{r}}$. ■

To compare the primal and dual results, note that despite the similarity of the bounds, the projection operators do not preserve the tight relationship between primal and dual updates. That is, even if $(1-\gamma)\mathbf{q} = H\mathbf{r}$ and $(1-\gamma)(\mathcal{O}\mathbf{q}) = (\mathcal{O}H)\mathbf{r}$, it is not true in general that $(1-\gamma)(\mathcal{PO}\mathbf{q}) = (\mathcal{PO}H)\mathbf{r}$. The most obvious difference comes from the fact that in the dual, the space of H matrices has bounded diameter, whereas in the primal, the space of $\mathbf{q}$ vectors has unbounded diameter in the natural norms. Automatically, the dual updates cannot diverge with compositions like $\mathcal{PO}$ and $\mathcal{PM}$.

5.2 Gradient Operator

In large scale problems one does not normally have the luxury of computing full dynamic programming updates that evaluate complete expectations over the entire domain, since this requires knowing the stationary visit distribution $\mathbf{z}$ for $P\Pi$ (essentially requiring one to know the model of the MDP). Moreover, full least squares projections are usually not practical to compute. A key intermediate step toward practical DP and RL algorithms is to formulate gradient step operators that only approximate full projections. Conveniently, the gradient update and projection operators are independent of the on-policy and off-policy updates and can be applied in either case. However, as we will see below, the gradient update operator causes significant instability in the off-policy update, to the degree that divergence is a common phenomenon (much more so than with full projections). Composing approximation with an off-policy update (max operator) in the primal case is very dangerous! All other operator combinations are much better behaved in practice, and even those that are not known to converge usually behave reasonably. Unfortunately, composing the gradient step with an off-policy update is a common algorithm attempted in reinforcement learning (Q-learning with function approximation), despite being the most unstable.

In the dual representation, one can derive a gradient update operator in a similar way to the primal, except that it is important to maintain the constraints on the parameters

$\mathbf{w}$, since the basis functions are probability distributions. We start by considering the projection objective

$$J_H \;=\; \frac{1}{2}\|H - \hat{H}\|_{\mathbf{z},\mathbf{r}}{}^2 \quad \text{subject to} \quad vec(\hat{H}) = \Psi\mathbf{w},\;\; \mathbf{w} \geq 0,\;\; \mathbf{w}^\top\mathbf{1} = \mathbf{1}$$

By the norm definition (see Equation 28), the above objective can be written as $J_H = \frac{1}{2}\|H\mathbf{r} - \hat{H}\mathbf{r}\|_{\mathbf{z}}{}^2$. Since $H\mathbf{r}$ and $\hat{H}\mathbf{r}$ are column vectors and $vec(\mathbf{AB}) = (\mathbf{B}^\top \otimes I)vec(\mathbf{A})$, we have $H\mathbf{r} = vec(H\mathbf{r}) = (\mathbf{r}^\top \otimes I)vec(H) = (\mathbf{r}^\top \otimes I)h$ and $\hat{H}\mathbf{r} = vec(\hat{H}\mathbf{r}) = (\mathbf{r}^\top \otimes I)vec(\hat{H}) = (\mathbf{r}^\top \otimes I)\hat{h}$, where $h \triangleq vec(H)$ and $\hat{h} \triangleq vec(\hat{H})$. Therefore, the objective becomes

$$J_H \;=\; \frac{1}{2}\|vec(H\mathbf{r}) - vec(\hat{H}\mathbf{r})\|_{\mathbf{z}}{}^2 \;=\; \frac{1}{2}\|(\mathbf{r}^\top \otimes I)(h - \Psi\mathbf{w})\|_{\mathbf{z}}{}^2$$

The unconstrained gradient of the above objective with respect to $\mathbf{w}$ is

$$\nabla_{\mathbf{w}} J_H \;=\; \Psi^\top(\mathbf{r}^\top \otimes I)^\top Z(\mathbf{r}^\top \otimes I)(\Psi\mathbf{w} - h) \;=\; \Gamma^\top Z(\mathbf{r}^\top \otimes I)(\hat{h} - h)$$

where $\Gamma = (\mathbf{r}^\top \otimes I)\Psi$, and $diag(Z)$ corresponds to the stationary distribution over state-action pairs. However, this gradient step cannot be followed directly because we need to maintain the constraints. The constraint $\mathbf{w}^\top\mathbf{1} = \mathbf{1}$ can be maintained by first projecting the gradient onto it, obtaining $\delta\mathbf{w} = (I - \frac{1}{k}\mathbf{1}\mathbf{1}^\top)\nabla_{\mathbf{w}} J_H$. Thus, the weight vector can be updated by

$$\mathbf{w}_{t+1} \;=\; \mathbf{w}_t - \alpha\delta\mathbf{w} \;=\; \mathbf{w}_t - \alpha(I - \frac{1}{k}\mathbf{1}\mathbf{1}^\top)\Gamma^\top Z(\mathbf{r}^\top \otimes I)(\hat{h} - h)$$

where α is a step-size parameter. Then the gradient operator can then be defined by

$$\mathcal{G}_{\hat{h}} h \;=\; \hat{h} - \alpha\Psi\delta\mathbf{w} \;=\; \hat{h} - \alpha\Psi(I - \frac{1}{k}\mathbf{1}\mathbf{1}^\top)\Gamma^\top Z(\mathbf{r}^\top \otimes I)(\hat{h} - h)$$

Since the target vector H (i.e., h) is determined by the underlying dynamic programming update, this gives the composed updates

$$\mathcal{G}_{\hat{h}}\mathcal{O}h = \alpha\Psi(I - \frac{1}{k}\mathbf{1}\mathbf{1}^\top)\Gamma^\top Z(\mathbf{r}^\top \otimes I)(\hat{h} - \mathcal{O}h) \;\text{ and }\; \mathcal{G}_{\hat{h}}\mathcal{M}h = \alpha\Psi(I - \frac{1}{k}\mathbf{1}\mathbf{1}^\top)\Gamma^\top(\mathbf{r}^\top \otimes I)(\hat{h} - \mathcal{M}h)$$

respectively for the on-policy and off-policy cases. Iterations that attempt to optimize $\hat{h}$ in the on-policy and off-policy cases respectively are given by

$$\hat{h}_{t+1} = \mathcal{G}_{\hat{h}_t}\mathcal{O}\hat{h}_t \qquad \text{and} \qquad \hat{h}_{t+1} = \mathcal{G}_{\hat{h}_t}\mathcal{M}\hat{h}_t$$

Thus far, the dual approach appears to hold an advantage over the standard primal approach, since convergence holds in every circumstance where the primal updates converge, and yet the dual updates are guaranteed never to diverge because the fundamental objects being represented are normalized probability distributions (i.e., belong to a bounded simplex). We now investigate the convergence properties of the various updates empirically.

6. Experimental Results

To investigate the effectiveness of the dual representations, we conducted experiments on various domains, including randomly synthesized MDPs, on the *star problem* (Baird, 1995), and on the *mountain car* problem. The randomly synthesized MDP domains allow us to test the general properties of the algorithms. The star problem is perhaps the most-cited example of a problem where Q-learning with linear function approximation diverges (Baird, 1995), and the mountain car domain has been prone to divergence with some primal representations as well (Boyan and Moore, 1995).

For each problem domain, twelve algorithms were run 100 times with a horizon of 1000 steps. The algorithms were: tabular on-policy ($\mathcal{O}$), projection on-policy ($\mathcal{PO}$), gradient on-policy ($\mathcal{GO}$), tabular off-policy ($\mathcal{M}$), projection off-policy ($\mathcal{PM}$), and gradient off-policy ($\mathcal{GM}$), for both the primal and the dual. The discount factor was set to $\gamma = 0.9$. For on-policy algorithms, we measure the difference between the values generated by the algorithms and those generated by the analytically determined fixed-point. For off-policy algorithms, we measure the difference between the values generated by the resulting policy and the values of the optimal policy. The step size for the gradient updates was 0.1 for primal representations and 100 for dual representations. The initial values of state-action value functions $\mathbf{q}$ are set according to the standard normal distribution, and state-action visit distributions H are chosen uniformly randomly with row normalization. Since the goal is to investigate the convergence of the algorithms without carefully crafting features, we also choose random basis functions according to a standard normal distribution for the primal representations, and random basis distributions according to a uniform distribution for the dual representations.

Randomly Synthesized MDPs. For the synthesized MDPs, we generated the transition and reward functions of the MDPs randomly—the transition function is uniformly distributed between 0 and 1 and the reward function is drawn from a standard normal. Here we only reported the results of random MDPs with 100 states, 5 actions, and 10 bases, but we observed consistent convergence of the dual representations on a variety of MDPs, with different numbers of states, actions, and bases. In Figure 2, the curve for the gradient off-policy update ($\mathcal{GM}$) in the primal case (solid line with the circle marker) blows up (diverges), while all the other curves (algorithms) in Figures 1 and 2 converge. Interestingly, the approximate error of the dual algorithm $\mathcal{PO}H$ (4.60×10^{-3}) is much smaller than the approximate error of the corresponding primal algorithm $\mathcal{PO}\mathbf{q}$ (4.23×10^{-2}), even though their theoretical bounds are the same (see Figure 1).

The Star Problem. The star problem has 7 states and 2 actions. The reward function is zero for each transition. In these experiments, we used the same fixed policy and linear value function approximation as in (Baird, 1995). In the dual, the number of bases is also set to 14 and the initial values of the state-action visit distribution matrix H are uniformly distributed random numbers between 0 and 1 with row normalization. The gradient off-policy update in the primal case diverges (see the solid line with the circle marker in Figure 4). However, all the updates with the dual representation algorithms converge.

The Mountain Car Problem The mountain car domain has continuous state and action spaces, which we discretized with a simple grid, resulting in an MDP with 222 states and 3

actions. The number of bases was chosen to be 5 for both the primal and dual algorithms. For the same reason as before, we chose the bases for the algorithms randomly. In the primal representations with linear function approximation, we randomly generated basis functions according to the standard normal distribution. In the dual representations, we randomly picked the basis distributions according to the uniform distribution. In Figure 6, we again observed divergence of the gradient off-policy update on state-action values in the primal, and the convergence of all the dual algorithms (see Figures 5 and 6). Again, the approximation error of the projected on-policy update $\mathcal{PO}H$ in the dual (1.90×10^1) is also considerably smaller than $\mathcal{PO}\mathbf{q}$ (3.26×10^2) in the primal (see Figure 5).

7. Conclusion

We investigated new dual representations for LP, DP and RL algorithms based on maintaining probability distributions, and explored connections to their primal counterparts based on maintaining value functions. In particular, we derived the original dual form representations from basic LP duality, extended these representations to derive new forms of DP algorithms and new forms of RL algorithms (TD evaluation, Sarsa, and Q-learning), and furthermore demonstrated how these dual representations can be scaled up to large domains by introducing normalized linear approximations. Although many of the results demonstrate equivalence between the primal and dual approaches, some advantages seem apparent for the dual approach, including an intrinsic robustness against divergence.

References

Leemon C. Baird. Residual algorithms: Reinforcement learning with function approximation. In *International Conference on Machine Learning*, pages 30–37, 1995.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, 1995.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Justin A. Boyan and Andrew W. Moore. Generalization in reinforcement learning: Safely approximating the value function. In *Neural Information Processing Systems*, pages 369–376, 1995.

Peter Dayan. Improving generalisation for temporal difference learning: The successor representation. *Neural Computation*, 5:613–624, 1993.

Daniela P. de Farias and Benjamin Van Roy. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal Optimization Theory and Applications*, 105(3):589–608, 2000.

Daniela P. de Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

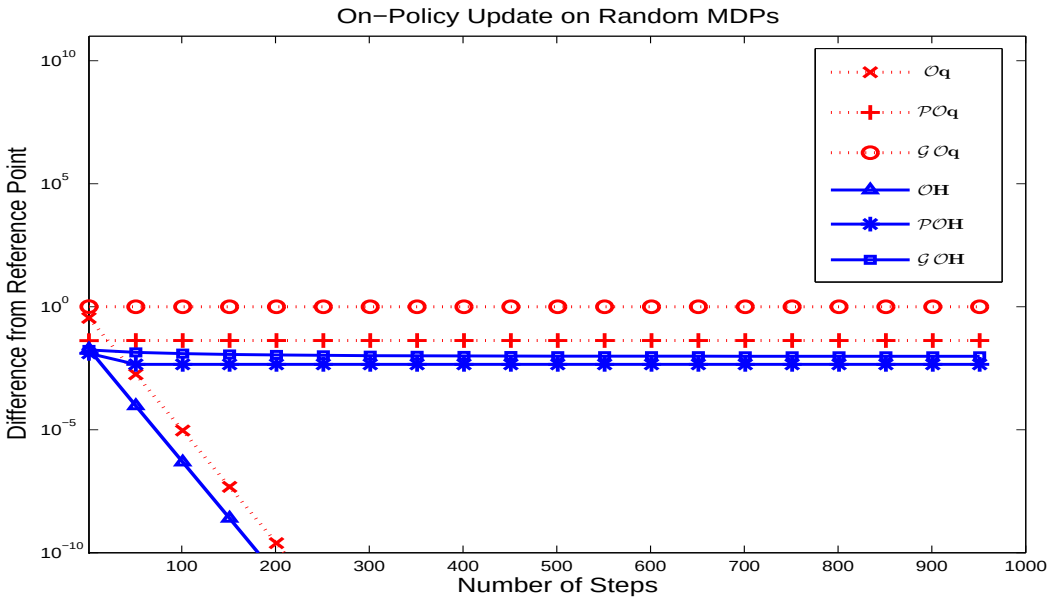

Figure 1: On-policy update of state-action value $\mathbf{q}$ and visit distribution H on randomly synthesized MDPs

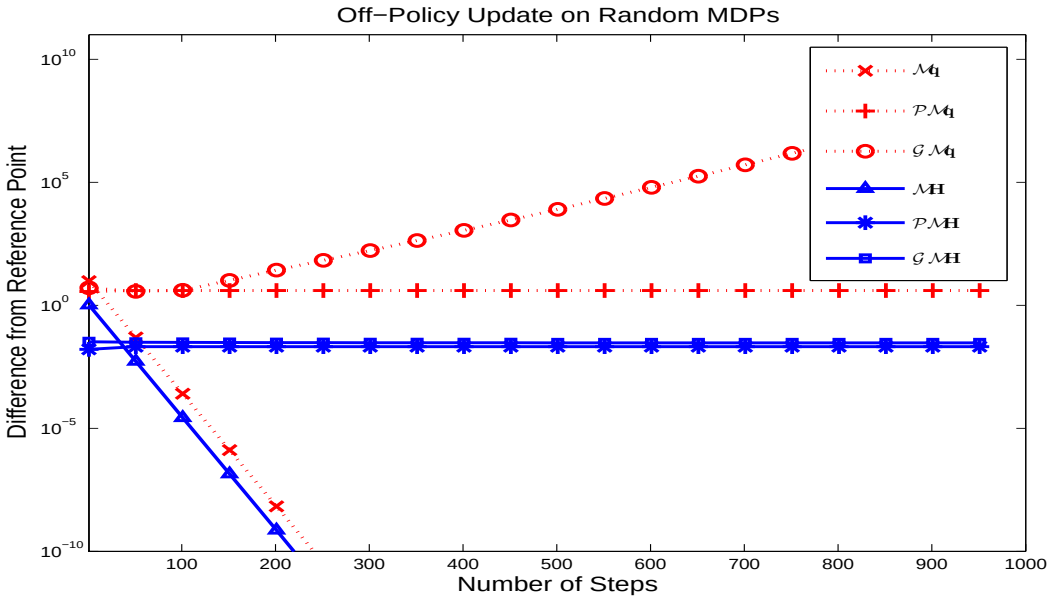

Figure 2: Off-policy update of state-action value $\mathbf{q}$ and visit distribution H on randomly synthesized MDPs

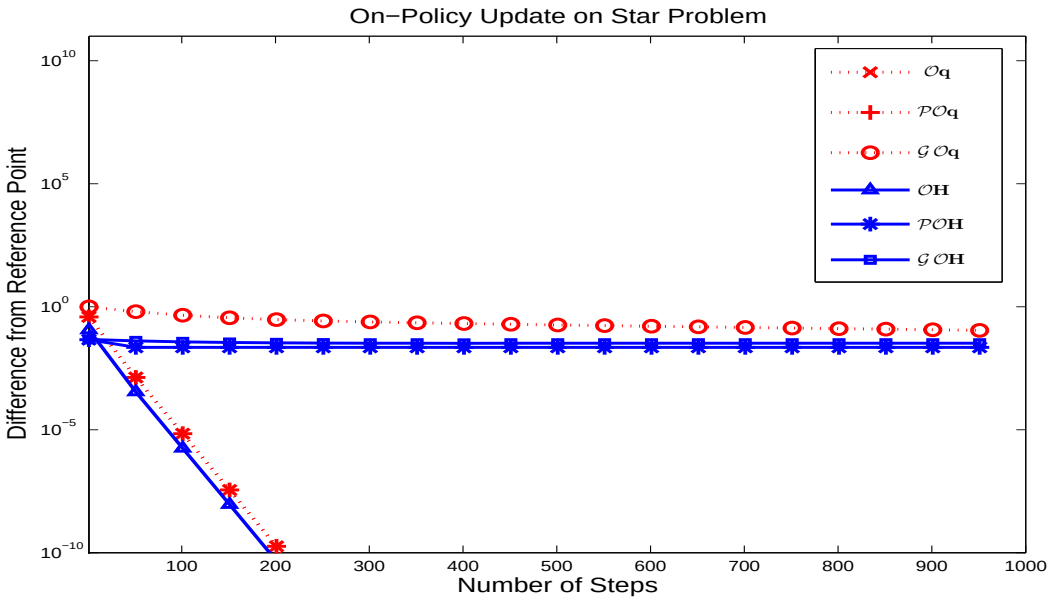

Figure 3: On-policy update of state-action value $\mathbf{q}$ and visit distribution H on the star problem

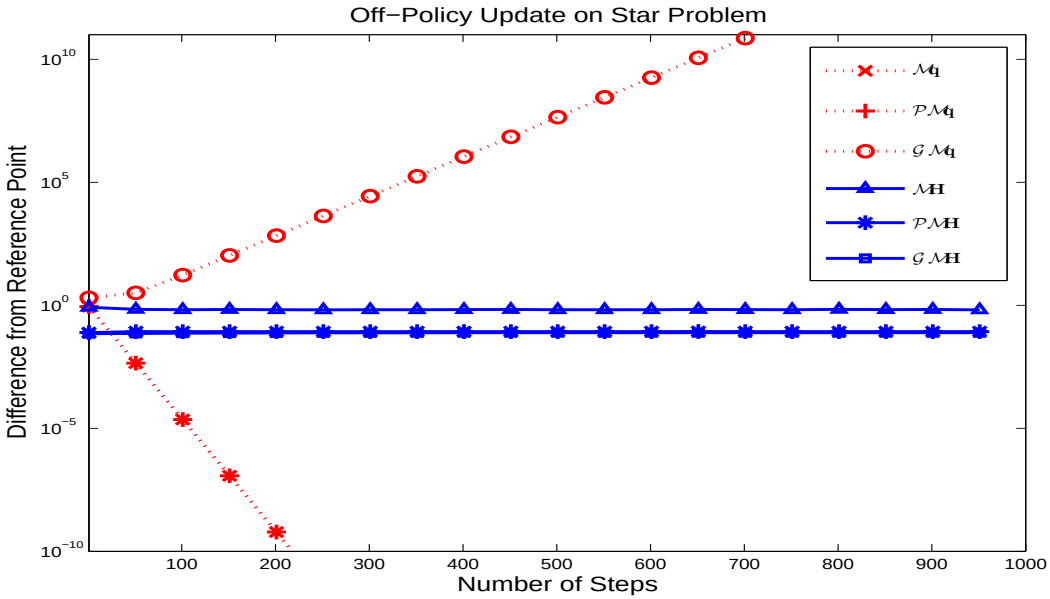

Figure 4: Off-policy update of state-action value $\mathbf{q}$ and visit distribution H on the star problem

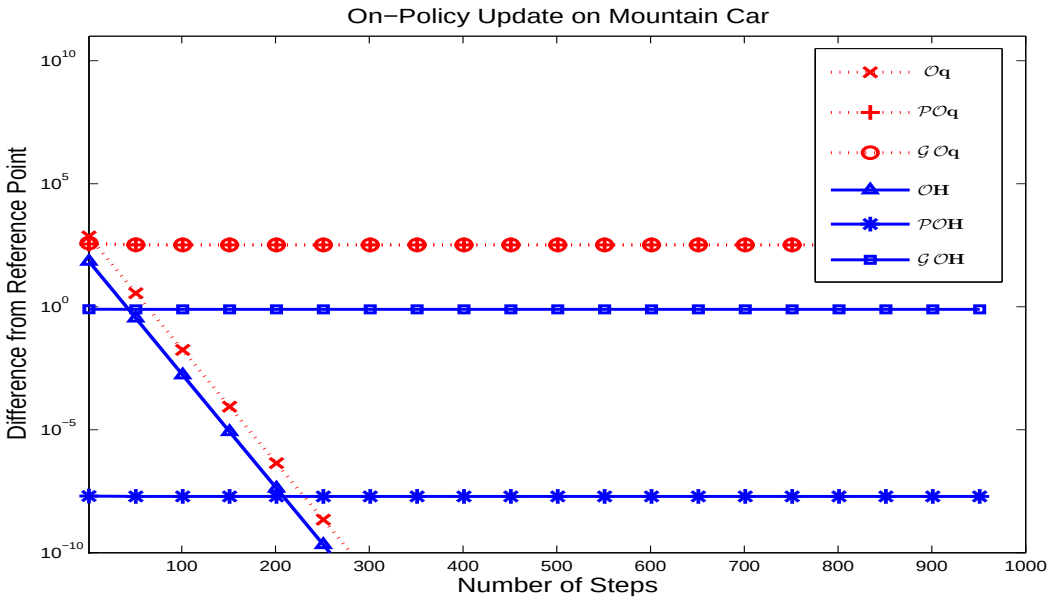

Figure 5: On-policy update of state-action value $\mathbf{q}$ and visit distribution H on the mountain car problem

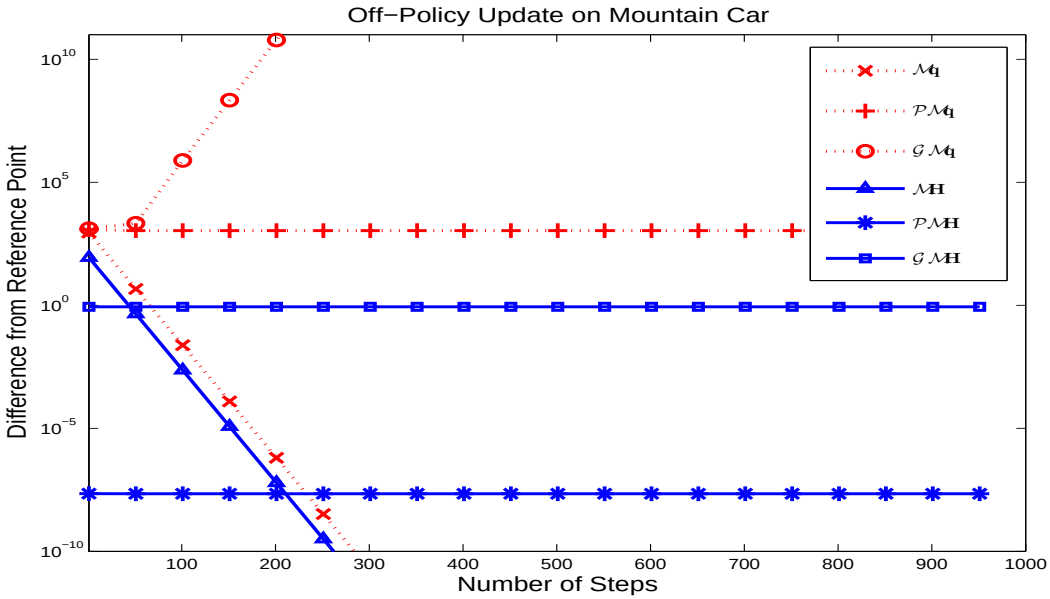

Figure 6: Off-policy update of state-action value $\mathbf{q}$ and visit distribution H on the mountain car problem

Andrew Y. Ng, Ronald Parr, and Daphne Koller. Policy search via density estimation. In *Neural Information Processing Systems*, 1999.

Martin L. Puterman. *Markov Decision Processes: Discrete Dynamic Programming*. Wiley, 1994.

Sheldon M. Ross. *Introduction to Probability Models*. Academic Press, 6th edition, 1997.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.