Journal of Machine Learning Research 1 (2008) 1-29 Submitted 1/08; Published ??/??

Dual Representations for Dynamic Programming

Tao Wang TAO.WANG@ANU.EDU.AU
Computer Sciences Laboratory
Australian National University
Canberra ACT 0200 Australia

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: Unknown

Abstract

We propose a dual approach to dynamic programming and reinforcement learning, based on maintaining an explicit representation of visit distributions as opposed to value functions. An advantage of working in the dual is that it allows one to exploit techniques for representing, approximating, and estimating probability distributions, while also avoiding any risk of divergence. We begin by formulating a modified dual of the standard linear program that guarantees the solution is a globally normalized visit distribution. Using this alternative representation, we then derive dual forms of dynamic programming, including on-policy updating, policy improvement and off-policy updating, and furthermore show how to incorporate function approximation. We then investigate the convergence properties of these algorithms, both theoretically and empirically, and show that the dual approach remains stable in situations when primal value function approximation diverges. Overall, the dual approach offers a viable alternative to standard dynamic programming techniques and offers new avenues for developing algorithms for sequential decision making.

Keywords: Sequential Decision Making, Dynamic Programming, Convergence, 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). In fact, the concept of value is so pervasive in DP and RL 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 and 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 distinct approach to solving DP and 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 the connection to LP duality. Nevertheless, as we will show, there exists a dual form for every standard value function algorithm, including on-policy updating, policy improvement, off-policy updating and variants using linear function approximation.

In this paper, we offer a systematic investigation of dual solution techniques based on manipulating state and state-action visit distributions instead of value functions, and moreover show how these techniques can be scaled up with linear function approximation. Beyond merely introducing dual DP algorithms, we also investigate their convergence properties. The proof techniques we use to analyze convergence are simple, but lead to useful conclusions. In particular, we find that the standard convergence results for value function based approaches also apply to the dual case, even in the presence of function approximation and off-policy updating. Although our results show that the dual approach yields equivalent results to the primal in the tabular case—as one would expect—the dual approach has an advantage when using approximation: since the fundamental objects being represented are normalized probability distributions (i.e., belong to a bounded simplex), dual updates cannot diverge. In particular, we find that dual updates usually converge in the very circumstance—gradient-based off-policy updates with linear function approximation (Baird, 1995; Sutton and Barto, 1998)—where primal updates can and often do diverge. Overall, we show that 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 stable DP and RL methods. This paper combines and extends the previous shorter contributions (Wang et al., 2007, 2008) and (Wang, 2007, Chapter 3).

2. Preliminaries

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

- 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 (that is, P is nonnegative and row normalized, where $P_{(sa,s')} = p(s' \mid s,a) \geq 0$ and $\sum_{s'} p(s' \mid s,a) = 1$ for all s,a); and

- 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., $\mathbf{r}_{(sa)} = \mathrm{E}[r \mid s,a]$.

We focus on 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 $0 \leq \gamma < 1$. In this case it is known that an optimal behavior strategy can always be expressed by a stationary policy. Initially, we will represent policies by $|S||A| \times 1$ vectors, $\boldsymbol{\pi}$, whose entries $\boldsymbol{\pi}_{(sa)}$ specify the probability of taking action a in state s; i.e., $\sum_a \boldsymbol{\pi}_{(sa)} = 1$ for all s. Stationarity refers to the fact that the action selection probabilities do not change over time. Beyond stationarity, it is also known that 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 central problem is to compute an optimal policy given either a complete specification of the environment, 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*"). We focus primarily on the planning problem in this paper.

3. Linear Programming

To establish a dual representation, we briefly review the LP approach for solving MDPs in the discounted infinite horizon case. Here we assume we are given the environment variables P and $\mathbf{r}$, the discount factor $0 \leq \gamma < 1$, and an 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)

$$\begin{aligned} & \min_{\mathbf{v}} (1-\gamma)\boldsymbol{\mu}^\top \mathbf{v} \quad \text{subject to} \quad \mathbf{v}_{(s)} \geq \mathbf{r}_{(sa)} + \gamma P_{(sa,:)}\mathbf{v} \quad \text{for all } s, a \\ = \; & \min_{\mathbf{v}} (1-\gamma)\boldsymbol{\mu}^\top \mathbf{v} \quad \text{subject to} \quad \Xi^\top \mathbf{v} \geq \mathbf{r} + \gamma P \mathbf{v} \qquad (1) \end{aligned}$$

where $\mathbf{v}$, an $|S| \times 1$ vector, is the state value function, and Ξ is an $|S| \times |S||A|$ matrix

$$\Xi = \begin{pmatrix} 1 \cdots 1 & & & \\ & 1 \cdots 1 & & \\ & & \ddots & \\ & & & 1 \cdots 1 \end{pmatrix}$$

given by $|S|$ row blocks of 1s, each of of length $|A|$, arranged block diagonally. The optimal solution to this LP corresponds to the *value function* for the optimal policy, from which the optimal policy can then be recovered by

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

Note that $\boldsymbol{\mu}$ and $(1-\gamma)$ behave as an arbitrary positive vector and positive constant in the LP and do not affect the minimizer, $\mathbf{v}^*$, 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, and we have chosen the objective in (1) in a specific way to obtain the following.

To derive our particular form of dual LP, consider an $|S||A| \times 1$ vector of Lagrange multipliers $\mathbf{d}$, and form the Lagrangian of (1) as

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

subject to $\mathbf{d} \geq 0$. Taking the gradient of the Lagrangian with respect to $\mathbf{v}$ and setting the result to zero yields $\Xi\mathbf{d} = (1-\gamma)\boldsymbol{\mu} + \gamma P^\top \mathbf{d}$. Substituting this constraint back into the Lagrangian eliminates $\mathbf{v}$ and yields the following dual LP

$$\max_{\mathbf{d}} \mathbf{d}^\top \mathbf{r} \quad \text{subject to} \quad \mathbf{d} \geq 0, \;\; \Xi\mathbf{d} = (1-\gamma)\boldsymbol{\mu} + \gamma P^\top \mathbf{d} \qquad (2)$$

Interestingly, 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. Let $\mathbf{1}$ denote the vector of all 1s.

Lemma 1 *The solution to (2) satisfies* $\mathbf{d} \geq 0$ *and* $\mathbf{1}^\top \mathbf{d} = 1$

Proof Nonnegativity is explicitly enforced by (2). Next, by the definition of Ξ, we have $\mathbf{1}^\top \mathbf{d} = \mathbf{1}^\top \Xi \mathbf{d}$. Then by (2), it follows that $\mathbf{1}^\top \mathbf{d} = (1-\gamma)\mathbf{1}^\top \boldsymbol{\mu} + \gamma \mathbf{1}^\top P^\top \mathbf{d}$. Since P is row normalized and $\boldsymbol{\mu}$ is a probability distribution, it follows that $\mathbf{1}^\top \mathbf{d} = (1-\gamma) + \gamma \mathbf{1}^\top \mathbf{d}$, which implies the result, provided $\gamma \neq 1$. ■

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 simple relation $\boldsymbol{\pi}^*_{(sa)} = \mathbf{d}^*_{(sa)} / \sum_a \mathbf{d}^*_{(sa)}$ (Ross, 1997). Note that the joint distribution $\mathbf{d}^*$ does *not* correspond to the stationary state-action visit distribution induced by $\boldsymbol{\pi}^*$ (unless $\gamma = 1$), but we will see below that it does correspond to a distribution of discounted state-action visits beginning in the initial state distribution $\boldsymbol{\mu}$. Thus, 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. We use this representation throughout the rest of the paper to derive novel forms of DP algorithms.

3.1 Policy Notation

Before we present dual algorithms and their convergence analysis, we find it convenient to re-express a policy $\boldsymbol{\pi}$ as an $|S| \times |S||A|$ matrix Π, where

$$\Pi = \begin{pmatrix} \mathbf{p}(a|s_1)^\top & & & \\ & \mathbf{p}(a|s_2)^\top & & \\ & & \ddots & \\ & & & \mathbf{p}(a|s_{|S|})^\top \end{pmatrix}$$

such that $\mathbf{p}(a|s_1)^\top = [\boldsymbol{\pi}_{(s_1 a_1)} \boldsymbol{\pi}_{(s_1 a_2)} \cdots \boldsymbol{\pi}_{(s_1 a_{|A|})}]$, a row vector. (Note that the same definition is also used in (Lagoudakis and Parr, 2003).) Although this representation of Π might appear unnatural, we find it 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 MDPs are typically expressed in terms of the primal value function. In this section, we focus on the tabular case and demonstrate that all classical DP algorithms have natural duals expressed in terms of state or state-action visit distributions. The DP algorithms are organized according to their update types: on-policy update, policy improvement, and off-policy update.

Note that the algorithms presented in this section are intended to be conceptual contributions—the new dual forms generally require more space than their primal counterparts, but lay the foundation for practical algorithms based on function approximation to be developed later in Section 6. Ultimately, we will see in Section 7 that the dual algorithms possess advantageous convergence properties.

4.1 On-Policy Update

First consider the straightforward problem of policy evaluation. Here we assume we are given a fixed policy Π and wish to compute either its value function in the primal or its discounted visit distribution in the dual.

4.1.1 State Based Policy Evaluation

Consider the simple case of state based policy evaluation.

Primal Representation. In the primal view, the role of policy evaluation is to recover the value function for a given policy Π, defined to be the expected sum of future discounted rewards. This definition can be compactly expressed in a vector-matrix form as

$$\mathbf{v} \;=\; \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi \mathbf{r} \tag{3}$$

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

$$\mathbf{v} \;=\; \Pi\mathbf{r} + \sum_{i=1}^{\infty} \gamma^i (\Pi P)^i \Pi \mathbf{r} \;=\; \Pi\mathbf{r} + \gamma(\Pi P) \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi \mathbf{r} \;=\; \Pi\mathbf{r} + \gamma \Pi P \mathbf{v} \tag{4}$$

Using this state value representation, a DP algorithm for policy evaluation can then be defined by repeatedly applying a vector operator $\mathcal{O}$, given by

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

For a given policy Π, repeated application of the on-policy operator $\mathcal{O}$ brings the current representation closer to satisfying the fixed point equation (4). (We examine the convergence properties of the on-policy operator in detail in Section 5.1 below, after the tabular DP algorithms have been introduced.)

Dual Vector Representation. To derive a *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

$$\mathbf{c}^\top \;=\; (1-\gamma)\boldsymbol{\mu}^\top \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \tag{5}$$

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

$$\mathbf{c}^\top \;=\; (1-\gamma)\boldsymbol{\mu}^\top + (1-\gamma)\boldsymbol{\mu}^\top \sum_{i=1}^{\infty} \gamma^i (\Pi P)^i \;=\; (1-\gamma)\boldsymbol{\mu}^\top + \gamma \mathbf{c}^\top \Pi P \tag{6}$$

It can be easily verified that (5) in fact defines a probability distribution over states.

Lemma 2 $\mathbf{c} \geq 0$ *and* $\mathbf{c}^\top \mathbf{1} = 1$

Proof First, $\mathbf{c} \geq 0$ holds since (5) is a convex combination of nonnegative terms. Second, it is easy to verify that (5) satisfies (6), implying that $\mathbf{c}^\top \mathbf{1} = (1-\gamma)\boldsymbol{\mu}^\top \mathbf{1} + \gamma \mathbf{c}^\top \Pi P \mathbf{1}$. Since ΠP is row normalized and $\boldsymbol{\mu}$ is a probability distribution, it follows that $\mathbf{c}^\top \mathbf{1} = (1-\gamma)1 + \gamma \mathbf{c}^\top \mathbf{1}$, which implies the result, provided $\gamma \neq 1$. ■

Not only is $\mathbf{c}$ a proper probability distribution over states, it also allows one to easily compute the expected discounted return of the policy Π.

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

Proof Plugging the definition of $\mathbf{v}$ (3) into the left of the above equation and then applying the definition of $\mathbf{c}$ (5) yields $(1-\gamma)\boldsymbol{\mu}^\top \mathbf{v} \;=\; (1-\gamma)\boldsymbol{\mu}^\top \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \Pi \mathbf{r} \;=\; \Big((1-\gamma)\boldsymbol{\mu}^\top \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i\Big) \Pi \mathbf{r} \;=\; \mathbf{c}^\top \Pi \mathbf{r}$. ■

Thus, a dual form of policy evaluation can be conducted by solving (6) for $\mathbf{c}$. The expected discounted reward obtained by policy Π 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 Π given only access to $\mathbf{c}$. Thus, we are compelled to extend the dual representation to a richer representation that avoids an implicit dependence on the specific initial distribution $\boldsymbol{\mu}$.

Dual Matrix Representation. Consider the following definition for an $|S| \times |S|$ matrix

$$M \;=\; (1-\gamma) \sum_{i=0}^{\infty} \gamma^i (\Pi P)^i \tag{7}$$

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

$$\begin{aligned} M \;&=\; (1-\gamma)I + (1-\gamma)\textstyle\sum_{i=1}^{\infty} \gamma^i (\Pi P)^i \;=\; (1-\gamma)I + \gamma M \Pi P \\ &\phantom{=\; (1-\gamma)I + (1-\gamma)\textstyle\sum_{i=1}^{\infty} \gamma^i (\Pi P)^i\;} =\; (1-\gamma)I + \gamma \Pi P M \end{aligned} \tag{8}$$

It can be easily verified that (7) defines a matrix where each row is a probability distribution.

Lemma 4 $M \geq 0$ *and* $M\mathbf{1} = \mathbf{1}$

Proof First, $M \geq 0$ holds since (7) is a convex combination of nonnegative terms. Second, it is easy to verify that (7) satisfies (8), implying that $M\mathbf{1} = (1-\gamma)\mathbf{1} + \gamma M \Pi P \mathbf{1}$. Since ΠP is row normalized, it follows that $M\mathbf{1} = (1-\gamma)\mathbf{1} + \gamma M\mathbf{1}$, which implies the result, provided $\gamma \neq 1$. ■

Furthermore, note that the matrix M shares a close relationship to $\mathbf{c}$, which can be seen more clearly by noting that each row of M satisfies

$$M = \begin{pmatrix} \mathbf{m}_1{}^\top \\ \mathbf{m}_2{}^\top \\ \vdots \\ \mathbf{m}_{|S|}{}^\top \end{pmatrix} = \begin{pmatrix} (1-\gamma)\mathbf{1}_1{}^\top + \gamma \mathbf{m}_1{}^\top \Pi P \\ (1-\gamma)\mathbf{1}_2{}^\top + \gamma \mathbf{m}_2{}^\top \Pi P \\ \vdots \\ (1-\gamma)\mathbf{1}_{|S|}{}^\top + \gamma \mathbf{m}_{|S|}{}^\top \Pi P \end{pmatrix}$$

where $\mathbf{1}_s$ is a vector of all zeros except for a 1 in the s^{th} position. That is, each row of M is a probability distribution (Lemma 4) where the entries $M_{(s,s')}$ correspond to the probability of discounted state visits to s' for a policy Π starting in state s—the same as $\mathbf{c}$ except with a different initial distribution. The matrix representation M has an advantage over the vector representation $\mathbf{c}$ by dropping the dependence on a particular $\boldsymbol{\mu}$, while still allowing $\mathbf{c}$ to be recovered by averaging M's rows according to $\boldsymbol{\mu}$.

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

Proof Starting from the definition of $\mathbf{c}$ (5) and applying the definition of M (7) yields $\mathbf{c}^\top = (1-\gamma)\boldsymbol{\mu}^\top \sum_{i=0}^\infty \gamma^i (\Pi P)^i = \boldsymbol{\mu}^\top \Big((1-\gamma)\sum_{i=0}^\infty \gamma^i (\Pi P)^i\Big) = \boldsymbol{\mu}^\top M$. ■

Lemmas 4 and 5 show that M is a variant of the "successor representation" proposed in (Dayan, 1993), but here extended to the infinite horizon discounted case. Importantly, not only is M a matrix of probability distributions over states, it allows one to easily recover the state values of the policy Π. That is, there exists an intimate connection between the primal state value function $\mathbf{v}$ and the dual state visit distribution M, which does not depend on the specific initial state distribution $\boldsymbol{\mu}$.

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

Proof Plugging the definition of $\mathbf{v}$ (3) into the left of the above equation and then applying the definition of M (7) yields $(1-\gamma)\mathbf{v} = (1-\gamma)\sum_{i=0}^\infty \gamma^i (\Pi P)^i \Pi \mathbf{r} = \Big((1-\gamma)\sum_{i=0}^\infty \gamma^i (\Pi P)^i\Big)\Pi\mathbf{r} = M\Pi\mathbf{r}$. ■

Thus, a dual form of policy evaluation can be conducted by solving (8) for M. Then at any time, an equivalent representation to $\mathbf{v}$ can be recovered by $M\Pi\mathbf{r}/(1-\gamma)$, as demonstrated by Theorem 6.

Note that there is a many to one relationship between the dual and primal representations respectively, because the number of variables in M exceeds the number of constraints relating M and $\mathbf{v}$ in Theorem 6. We will obtain a more compact representation by incorporating function approximation in Section 6 below.

Dual Update Operator. Even though the above shows that policy evaluation can be performed by solving a system of linear equations, the linear system also provides an online update operator that can be repeatedly applied to reach the solution, as in the primal case. That is, a DP algorithm for policy evaluation can be defined by repeatedly applying a matrix operator $\mathcal{O}$, given by

$$\mathcal{O}M \quad = \quad (1-\gamma)I + \gamma \Pi P M \tag{9}$$

For a given policy Π—as we will discuss in Section 5.1 below—repeated application of the on-policy operator $\mathcal{O}$ brings the current representation closer to satisfying the fixed point equation (8).

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, to formulate dual variants of classical RL algorithms such as Sarsa and Q-learning, we will ultimately need to use *state-action* based evaluations. In this section we derive the state-action analogues to the previous state based algorithms.

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

$$\mathbf{q} \quad = \quad \sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \mathbf{r} \tag{10}$$

This state-action value function is closely related to the previous state value function (3) and satisfies a similar recursive relation

$$\mathbf{q} \quad = \quad \mathbf{r} + \gamma P \Pi \mathbf{q} \tag{11}$$

As in the state based case, a DP algorithm for policy evaluation can be defined by repeatedly applying a vector operator $\mathcal{O}$, given by

$$\mathcal{O}\mathbf{q} \quad = \quad \mathbf{r} + \gamma P \Pi \mathbf{q} \tag{12}$$

For a given policy Π—as we discuss formally in Section 5.1 below—repeated application of the on-policy operator brings the current representation closer to satisfying the fixed point equation (11).

Dual Vector Representation. To develop a dual form of state-action policy evaluation, we represent a probability distribution over state-action pairs that has a useful correspondence to the long run expected discounted rewards obtained by the policy. Such a correspondence can be achieved by defining the following probability distribution over state-action pairs

$$\mathbf{d}^\top \quad = \quad (1-\gamma)\boldsymbol{\nu}^\top \sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \tag{13}$$

where $\boldsymbol{\nu}$ is an $|S||A| \times 1$ vector specifying the initial distribution over state-action pairs given by $\boldsymbol{\nu} = \Pi^\top \boldsymbol{\mu}$.

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

$$\mathbf{d}^\top \;\; = \;\; (1-\gamma)\boldsymbol{\nu}^\top + \gamma \mathbf{d}^\top P \Pi \tag{14}$$

It can be easily verified that (13) defines a probability distribution over state-action pairs.

Lemma 7 $\mathbf{d} \geq 0$ *and* $\mathbf{d}^\top \mathbf{1} = 1$

Proof First, $\mathbf{d} \geq 0$ holds since (13) is a convex combination of nonnegative terms. Second, it is easy to verify that (13) satisfies (14), implying that $\mathbf{d}^\top \mathbf{1} = (1-\gamma)\boldsymbol{\nu}^\top \mathbf{1} + \gamma \mathbf{d}^\top P \Pi \mathbf{1}$. Since $P\Pi$ is row normalized and $\boldsymbol{\nu}$ is a probability distribution, it follows that $\mathbf{d}^\top \mathbf{1} = (1-\gamma)1 + \gamma \mathbf{d}^\top \mathbf{1}$, which implies the result, provided $\gamma \neq 1$. ■

Interestingly, not only is $\mathbf{d}$ a proper probability distribution over state-action pairs, it also is automatically guaranteed to be a feasible point in the dual LP (2).

Lemma 8 $\mathbf{d}$ *is a feasible point for (2)*

Proof First, it is obvious that $\mathbf{d} \geq 0$, since (13) is a convex combination nonnegative terms. To verify that the equality constraint in (2) is satisfied, note that again it is easy to verify (13) satisfies (14). Thus $\Xi \mathbf{d} = (1-\gamma)\Xi\boldsymbol{\nu} + \gamma \Xi \Pi^\top P^\top \mathbf{d} = (1-\gamma)\boldsymbol{\mu} + \gamma P^\top \mathbf{d}$, using the facts that $\Xi\Pi^\top = I$ and $\Xi\boldsymbol{\nu} = \Xi\Pi^\top\boldsymbol{\mu} = \boldsymbol{\mu}$, which gives the result. ■

Therefore, this form of $\mathbf{d}$ constitutes a valid dual representation, which can be moreover used to recover the expected discounted return of the policy Π.

Lemma 9 $(1-\gamma)\boldsymbol{\nu}^\top \mathbf{q} = \mathbf{d}^\top \mathbf{r}$

Proof Plugging the definition of $\mathbf{q}$ (10) into the left of the above equation and then applying the definition of $\mathbf{d}$ (13) yields $(1-\gamma)\boldsymbol{\nu}^\top \mathbf{q} \;=\; (1-\gamma)\boldsymbol{\nu}^\top \sum_{i=0}^\infty \gamma^i (P\Pi)^i \mathbf{r} \;=\; \Big((1-\gamma)\boldsymbol{\nu}^\top \sum_{i=0}^\infty \gamma^i (P\Pi)^i\Big)\mathbf{r} \;=\; \mathbf{d}^\top \mathbf{r}$. ■

These results show that a dual form of state-action policy evaluation can be conducted by solving (14) for $\mathbf{d}$. The expected discounted reward obtained by policy Π starting in the initial state-action distribution given by $\boldsymbol{\nu}^\top = \boldsymbol{\mu}^\top \Pi$ can then be computed by $\mathbf{d}^\top \mathbf{r}/(1-\gamma)$, according to Lemma 9. In principle, this gives another valid form of policy evaluation in a dual representation.

However, once again we will find that merely recovering the state-action distribution $\mathbf{d}$ is inadequate for *policy improvement* (Section 4.2 below), since there is no apparent way to improve Π 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 $\boldsymbol{\nu}$.

Dual Matrix Representation. Consider the following definition for an $|S||A| \times |S||A|$ matrix

$$H \;=\; (1-\gamma)\sum_{i=0}^{\infty}\gamma^i (P\Pi)^i \tag{15}$$

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

$$\begin{aligned} H \;&=\; (1-\gamma)I + (1-\gamma)\textstyle\sum_{i=1}^{\infty}\gamma^i (P\Pi)^i \;&=&\; (1-\gamma)I + \gamma H P\Pi \\ & &=&\; (1-\gamma)I + \gamma P\Pi H \end{aligned} \tag{16}$$

It can be easily verified that (15) defines a matrix where each row is a probability distribution.

Lemma 10 *$H \geq 0$ and $H\mathbf{1} = \mathbf{1}$*

Proof First, $H \geq 0$ holds since (15) is a convex combination of nonnegative terms. Second, it is easy to verify that (15) satisfies (16), implying that $H\mathbf{1} = (1-\gamma)\mathbf{1} + \gamma H P\Pi\mathbf{1}$. Since $P\Pi$ is row normalized, it follows that $H\mathbf{1} = (1-\gamma)\mathbf{1} + \gamma H\mathbf{1}$, which implies the result, provided $\gamma \neq 1$. ■

Once again, we find that the matrix H shares a strong relationship with its vector correspondent $\mathbf{d}$, since each row of H satisfies

$$H = \begin{pmatrix} \mathbf{h}_1{}^\top \\ \mathbf{h}_2{}^\top \\ \vdots \\ \mathbf{h}_{|S|}{}^\top \end{pmatrix} = \begin{pmatrix} (1-\gamma)\mathbf{1}_1{}^\top + \gamma\mathbf{h}_1{}^\top P\Pi \\ (1-\gamma)\mathbf{1}_2{}^\top + \gamma\mathbf{h}_2{}^\top P\Pi \\ \vdots \\ (1-\gamma)\mathbf{1}_{|S|}{}^\top + \gamma\mathbf{h}_{|S|}{}^\top P\Pi \end{pmatrix}$$

That is, each row of H is a probability distribution (Lemma 10) where the entries $H_{(sa,s'a')}$ correspond to the probability of discounted state-action visits to $(s'a')$ for a policy Π starting in state-action pair (sa)—the same as $\mathbf{d}$ except with a different initial distribution. The matrix representation H has an advantage over the vector representation $\mathbf{d}$ by dropping the dependence on a particular $\boldsymbol{\nu}$, while still allowing $\mathbf{d}$ to be recovered by averaging H's rows according to $\boldsymbol{\nu}$.

Lemma 11 $\mathbf{d}^\top \;=\; \boldsymbol{\nu}^\top H$

Proof Starting from the definition of $\mathbf{d}$ (13) and applying the definition of H (15) yields $\mathbf{d}^\top \;=\; (1-\gamma)\boldsymbol{\nu}^\top \sum_{i=0}^{\infty}\gamma^i(P\Pi)^i \;=\; \boldsymbol{\nu}^\top\Big((1-\gamma)\sum_{i=0}^{\infty}\gamma^i(P\Pi)^i\Big) \;=\; \boldsymbol{\nu}^\top H.$ ■

Not only is H a matrix of probability distributions over state-action pairs, it also allows one to easily recover the state-action values of the policy Π. That is, there also exists an important connection between the primal state-action value function $\mathbf{q}$ and the dual state-action visit distribution H, which does not depend on the specific initial state-action distribution $\boldsymbol{\nu}$.

Theorem 12 $(1-\gamma)\mathbf{q} = H\mathbf{r}$

Proof Plugging the definition of $\mathbf{q}$ (10) into the left of the above theorem and then applying the definition of H (15) yields $(1-\gamma)\mathbf{q} = (1-\gamma)\sum_{i=0}^{\infty}\gamma^i(P\Pi)^i\mathbf{r} = \Big((1-\gamma)\sum_{i=0}^{\infty}\gamma^i(P\Pi)^i\Big)\mathbf{r} = H\mathbf{r}$. ■

These results show that a dual form of state-action policy evaluation can be conducted by solving (16) for H. Then at any time, an equivalent representation to $\mathbf{q}$ can be recovered by $H\mathbf{r}/(1-\gamma)$, as established by Theorem 12.

Note however that there is a many to one relationship between the dual and primal representations, respectively, because the number of variables in H exceeds the number of constraints relating H and $\mathbf{q}$ in Theorem 12. As before, we will obtain a more compact representation by incorporating function approximation in Section 6 below.

Dual Update Operator. Now, as in the state based case, a DP algorithm for policy evaluation can be defined by repeatedly applying a matrix operator $\mathcal{O}$, given by

$$\mathcal{O}H = (1-\gamma)I + \gamma P\Pi H \tag{17}$$

For a given policy Π—as we will prove in Section 5.1 below—repeated application of the on-policy operator $\mathcal{O}$ brings the current representation closer to satisfying the fixed point equation (16).

4.1.3 State Versus State-Action Based Representations

The state and state-action based representations are closely related. In the primal case, the state value vector $\mathbf{v}$ and state-action value vector $\mathbf{q}$ are related by

Lemma 13 $\mathbf{v} = \Pi\mathbf{q}$

Proof Starting from the definition of $\mathbf{v}$ (3) and applying the definition of $\mathbf{q}$ (10) yields $\mathbf{v} = \sum_{i=0}^{\infty}\gamma^i(\Pi P)^i\Pi\mathbf{r} = \Pi\sum_{i=0}^{\infty}\gamma^i(P\Pi)^i\mathbf{r} = \Pi\mathbf{q}$. ■

In the dual case, the state visit matrix M and state-action visit matrix H are related by

Lemma 14 $M\Pi = \Pi H$

Proof Starting from the definition of M (7) and applying the definition of H (15) yields $M\Pi = (1-\gamma)\sum_{i=0}^{\infty}\gamma^i(\Pi P)^i\Pi = (1-\gamma)\sum_{i=0}^{\infty}\gamma^i\Pi(\Pi P)^i = \Pi(1-\gamma)\sum_{i=0}^{\infty}\gamma^i(\Pi P)^i = \Pi H$. ■

Given this close relationship, we will concentrate mainly on the state-action based representations below. Most of the results for the state-action based case immediately apply to the state based case.

To this point, we have developed new dual representations that can form the basis for both state based and state-action based policy *evaluation*. The alternative dual DP algorithms are defined in terms of state distributions and state-action distributions, and do not require value functions to be computed. Before we examine the convergence properties of these on-policy updates in detail, however, we first address policy improvement and introduce off-policy updates.

4.2 Policy Improvement

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

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

$$\boldsymbol{\pi}'_{(sa)} \;=\; \begin{cases} 1 & \text{if } a = a'(s) \\ 0 & \text{if } a \neq a'(s) \end{cases} \quad \text{such that} \quad \begin{aligned} a'(s) &= \arg\max\nolimits_a \mathbf{q}_{(sa)} \\ &= \arg\max\nolimits_a \mathbf{r}_{(sa)} + \gamma P_{(sa,:)}\mathbf{v} \end{aligned} \tag{18}$$

If we let Π' denote the matrix representation of the updated policy, then (18) yields the inequality $\Pi'\mathbf{q} \geq \Pi\mathbf{q}$ by construction. The subsequent "policy improvement theorem" (Sutton and Barto, 1998) verifies that this update leads to an improved policy.

Theorem 15 $\Pi\mathbf{q} \leq \Pi'\mathbf{q}$ *implies* $\mathbf{v} \leq \mathbf{v}'$

Proof It is instructive to briefly illustrate the result by expansion. First note that $\mathbf{q} = \mathbf{r} + \gamma P\mathbf{v}$, by (11) and Lemma 13. Combining this with the assumption $\Pi\mathbf{q} \leq \Pi'\mathbf{q}$ yields $\Pi(\mathbf{r} + \gamma P\mathbf{v}) \leq \Pi'(\mathbf{r} + \gamma P\mathbf{v})$. Finally, using (4) establishes the key inequality $\mathbf{v} = \Pi(\mathbf{r} + \gamma P\mathbf{v}) \leq \Pi'\mathbf{r} + \gamma\Pi'P\mathbf{v}$. This fact then immediately yields the chain of inequalities

$$\begin{aligned} \mathbf{v} &\leq \Pi'\mathbf{r} + \gamma\Pi'P\mathbf{v} \\ &\leq \Pi'\mathbf{r} + \gamma\Pi'P\Pi'\mathbf{r} + \gamma^2(\Pi'P)^2\mathbf{v} \\ &\leq \Pi'\mathbf{r} + \gamma\Pi'P\Pi'\mathbf{r} + \gamma^2(\Pi'P)^2\Pi'\mathbf{r} + \gamma^3(\Pi'P)^3\mathbf{v} \\ &\vdots \\ &\leq \sum_{i=0}^{\infty}\gamma^i(\Pi'P)^i\Pi'\mathbf{r} \;=\; \mathbf{v}' \end{aligned}$$

■

Dual Representation. The above development can be paralleled 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 representation one can derive an improved policy $\boldsymbol{\pi}'$ by the update

$$\boldsymbol{\pi}'_{(sa)} = \begin{cases} 1 & \text{if } a = a'(s) \\ 0 & \text{if } a \neq a'(s) \end{cases} \quad \text{such that} \quad \begin{aligned} a'(s) &= \arg\max\nolimits_a H_{(sa,:)}\mathbf{r} \\ &= \arg\max\nolimits_a (1-\gamma)\mathbf{r}_{(sa)} + \gamma P_{(sa,:)}M\Pi\mathbf{r} \end{aligned} \tag{19}$$

If we let Π' denote the matrix representation of the updated policy, then (19) yields the inequality $\Pi'H\mathbf{r} \geq \Pi H\mathbf{r}$ by construction. In fact, by Theorem 12, the two policy updates given in (18) and (19) respectively, must lead to the same resulting policy Π'. Therefore, not surprisingly, we have an analogous policy improvement theorem in this case.

Theorem 16 $\Pi H\mathbf{r} \leq \Pi' H\mathbf{r}$ *implies* $M\Pi\mathbf{r} \leq M'\Pi'\mathbf{r}$

Proof A formal proof proceeds by induction, but it is more instructive to illustrate the result by expansion. First, note that $H = (1-\gamma)I + \gamma PM\Pi$ by (16) and Lemma 14. Combining this with the assumption $\Pi H\mathbf{r} \leq \Pi' H\mathbf{r}$ yields $(1-\gamma)\Pi\mathbf{r} + \gamma\Pi PM\Pi\mathbf{r} \leq (1-\gamma)\Pi'\mathbf{r} + \gamma\Pi' PM\Pi\mathbf{r}$. Finally, (8) establishes the key inequality $M\Pi\mathbf{r} = ((1-\gamma)I + \gamma\Pi PM)\,\Pi\mathbf{r} \leq (1-\gamma)\Pi'\mathbf{r} + \gamma\Pi' P(M\Pi\mathbf{r})$. This fact can then be used to derive the chain of inequalities

$$\begin{aligned}
M\Pi\mathbf{r} &\leq (1-\gamma)\Pi'\mathbf{r} + \gamma\Pi' P(M\Pi\mathbf{r}) \\
&\leq (1-\gamma)\Pi'\mathbf{r} + (1-\gamma)\gamma\Pi' P\Pi'\mathbf{r} + \gamma^2(\Pi' P)^2(M\Pi\mathbf{r}) \\
&\leq (1-\gamma)\Pi'\mathbf{r} + (1-\gamma)\gamma\Pi' P\Pi'\mathbf{r} + (1-\gamma)\gamma^2(\Pi' P)^2\Pi'\mathbf{r} + \gamma^3(\Pi' P)^3(M\Pi\mathbf{r}) \\
&\vdots \\
&\leq (1-\gamma)\sum_{i=0}^{\infty}\gamma^i(\Pi' P)^i\Pi'\mathbf{r} \;=\; M'\Pi'\mathbf{r}
\end{aligned}$$

■

Therefore, a policy iteration algorithm can be completely expressed in terms of the dual representation, incorporating both dual policy evaluation and dual policy improvement (19). The resulting approach is equivalent to the standard primal policy iteration algorithms using primal policy improvement (18). In particular, each policy evaluation and policy improvement step maintains an equivalence between the primal and dual representations in the tabular case. (This equivalence will no longer be maintained when we consider approximate algorithms in Section 6.)

4.3 Off-Policy Update

Finally, beyond policy evaluation and policy iteration, *off-policy* updating provides a prominent basis for DP and RL algorithms; leading, for example, to value iteration and Q-learning algorithms. Off-policy updating is based on an alternative operator, $\mathcal{M}$, distinct from the on-policy operator $\mathcal{O}$, in that it is neither linear nor defined by a reference policy. Instead $\mathcal{M}$ employs a greedy *maximum* update to the current estimates.

Primal Representation. In the primal case, off-policy updates correspond to the standard value iteration algorithms. In the state based representation, the off-policy operator $\mathcal{M}$ is given by

$$\mathcal{M}\mathbf{v} \;=\; \Pi^*[\mathbf{r} + \gamma P\mathbf{v}] \quad \text{where} \quad \Pi^*[\mathbf{r} + \gamma P\mathbf{v}]_{(s)} \;=\; \max_a \mathbf{r}_{(sa)} + \gamma P_{(sa,:)}\mathbf{v} \tag{20}$$

The goal of this greedy update is to bring the representation $\mathbf{v}$ closer to satisfying the Bellman equation $\mathbf{v} = \Pi^*[\mathbf{r} + \gamma P\mathbf{v}]$. Similarly, in the state-action based representation, the off-policy operator $\mathcal{M}$ is given by

$$\mathcal{M}\mathbf{q} \;=\; \mathbf{r} + \gamma P\Pi^*[\mathbf{q}] \quad \text{where} \quad \Pi^*[\mathbf{q}]_{(s)} \;=\; \max_a \mathbf{q}_{(sa)} \tag{21}$$

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

Dual Representation. In the dual representation, analogues to the primal off-policy DP can be derived without any explicit representation of value. For succinctness, we focus on the state-action case only. Here, an off-policy update for the state-action visit distribution H can be expressed by

$$\begin{aligned} \mathcal{M}H &= (1-\gamma)I + \gamma P\Pi^*_{\mathbf{r}}[H], && \text{where} \\ \Pi^*_{\mathbf{r}}[H]_{(s,:)} &= H_{(sa'(s),:)} && \text{such that} \\ a'(s) &= \arg\max_a [H\mathbf{r}]_{(sa)} &= \arg\max_a \sum_{(s'a')} H_{(sa,s'a')}\mathbf{r}_{(s'a')} \end{aligned} \tag{22}$$

The goal of this greedy update is to bring the representation H closer to satisfying the Bellman equation $H = (1-\gamma)I + \gamma P\Pi^*_{\mathbf{r}}[H]$. Note that this off-policy dual DP algorithm does not refer to the primal value function at all. The convergence of this operator is established in Section 5.2 below.

Overall, we have developed a series of novel DP algorithms based on an alternative but fully expressive representation: normalized state and state-action visit distributions. These algorithms do not require value functions to be explicitly computed. In the tabular case, the dual algorithms appear to require more space than their primal counterparts—a shortcoming that will be rectified when we consider function approximation in Section 6. However, first, we examine the theoretical convergence properties of the basic tabular algorithms.

5. Convergence Analysis

We establish that the DP operators on the dual representations exhibit the same convergence properties as their primal counterparts in the tabular case. To keep the presentation succinct, we will concentrate only on state-action based representations, $\mathbf{q}$ and H—analogous results are easily obtained for the state-based representations, $\mathbf{v}$ and M.

5.1 On-policy Convergence

For the on-policy operator $\mathcal{O}$, convergence to a fixed point is proved by establishing a contraction property with respect to a specific norm (Tsitsiklis and Van Roy, 1997). In particular, one defines a weighted 2-norm where the weights are determined by the stationary distribution of the policy Π and transition model P. Let $\mathbf{z} \geq 0$ be a vector such that

$$\mathbf{z}^\top P\Pi = \mathbf{z}^\top \tag{23}$$

That is, $\mathbf{z}$ is the stationary state-action visit distribution for $P\Pi$. Note that $\mathbf{z}$ is not the same as the initial distribution $\boldsymbol{\nu}$ nor the discounted stationary distribution $\mathbf{d}$. Let $Z = \text{diag}(\mathbf{z})$.

Primal Representation. The steps taken in the primal case are useful to proving convergence in the dual. To establish convergence in the primal, one first defines a weighted 2-norm on $\mathbf{q}$ vectors

$$\|\mathbf{q}\|_{\mathbf{z}}{}^2 = \mathbf{q}^\top Z\mathbf{q} = \sum_{(sa)} \mathbf{z}_{(sa)}\mathbf{q}^2_{(sa)} \tag{24}$$

Crucially, for this norm, a state-action transition is a non-expansion; that is, it can be shown that $\|P\Pi\mathbf{q}\|_\mathbf{z} \leq \|\mathbf{q}\|_\mathbf{z}$ (Tsitsiklis and Van Roy, 1997). This fact can then be used to show that the on-policy operator $\mathcal{O}$ is a contraction: $\|\mathcal{O}\mathbf{q}_1 - \mathcal{O}\mathbf{q}_2\|_\mathbf{z} \leq \gamma\|\mathbf{q}_1 - \mathbf{q}_2\|_\mathbf{z}$ (Tsitsiklis and Van Roy, 1997). Finally, by the contraction map fixed point theorem (Bertsekas, 1995) there must exist a unique fixed point of $\mathcal{O}$ in the space of vectors $\mathbf{q}$—that is, a vector $\mathbf{q}_\Pi$ where $\mathbf{q}_\Pi = \mathcal{O}\mathbf{q}_\Pi$—such that repeated applications of $\mathcal{O}$ converges to $\mathbf{q}_\Pi$.

Dual Representation. 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. To do so, first define the pseudo-norm

$$\|H\|_{\mathbf{z},\mathbf{r}}{}^2 \;=\; \|H\mathbf{r}\|_\mathbf{z}{}^2 \;=\; \sum_{(sa)} \mathbf{z}_{(sa)} \Big(\sum_{(s'a')} H_{(sa,s'a')}\mathbf{r}_{(s'a')} \Big)^2 \tag{25}$$

It is easily verified that this definition satisfies the pseudo-norm properties; in particular, it satisfies the triangle inequality. Note that this definition depends on the stationary distribution $\mathbf{z}$ and the reward vector $\mathbf{r}$, hence the magnitude of a row normalized matrix H is determined by the magnitude of the weighted reward expectations it induces. Interestingly, this definition allows one to establish the same non-expansion and contraction results as the primal case. First, state-action transitions remain a non-expansion.

Lemma 17 $\|P\Pi H\|_{\mathbf{z},\mathbf{r}} \;\leq\; \|H\|_{\mathbf{z},\mathbf{r}}$

Proof From Jensen's inequality, we obtain

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

From the definition (25) it follows that $\|P\Pi H\|_{\mathbf{z},\mathbf{r}} = \|P\Pi(H\mathbf{r})\|_\mathbf{z} \leq \|H\mathbf{r}\|_\mathbf{z} = \|H\|_{\mathbf{z},\mathbf{r}}$. ■

This non-expansion result can be used to prove that the on-policy operator $\mathcal{O}$ is a contraction with respect to $\|\cdot\|_{\mathbf{z},\mathbf{r}}$.

Lemma 18 $\|\mathcal{O}H_1 - \mathcal{O}H_2\|_{\mathbf{z},\mathbf{r}} \;\leq\; \gamma\|H_1 - H_2\|_{\mathbf{z},\mathbf{r}}$

Proof From 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 17, this establishes $\|\mathcal{O}H_1 - \mathcal{O}H_2\|_{\mathbf{z},\mathbf{r}} \leq \gamma\|H_1 - H_2\|_{\mathbf{z},\mathbf{r}}$. ■

Therefore, once again by the contraction map fixed point theorem (Bertsekas, 1995) there exists a fixed point of $\mathcal{O}$ among row normalized matrices H, such that repeated applications of $\mathcal{O}$ will converge to matrices H_Π where $\mathcal{O}H_\Pi = H_\Pi$. 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 fixed point equation $\mathcal{O}H_\Pi = H_\Pi$ only specifies $|S||A|$ linear constraints on $|S|^2|A|^2$ unknowns. The set of fixed points forms a convex subspace (in fact, a simplex), since if $H_1\mathbf{r} = (1-\gamma)I + \gamma P\Pi H_1$ and $H_2\mathbf{r} = (1-\gamma)I + \gamma P\Pi H_2$ then $(\alpha H_1 + (1-\alpha)H_2)\mathbf{r} = (1-\gamma)I + \gamma P\Pi(\alpha H_1 + (1-\alpha)H_2)$ for any α, where furthermore α must be restricted to $0 \le \alpha \le 1$ to maintain nonnegativity. Hence, the on-policy operator converges to a simplex of equivalent fixed points $\{H : \mathcal{O}H = H\}$.

5.2 Off-policy Convergence

The strategy for establishing convergence for the off-policy operator $\mathcal{M}$ 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.

Primal Representation. In the primal representation, the max-norm is given by

$$\|\mathbf{q}\|_\infty \quad = \quad \max_{(sa)} |q_{(sa)}| \tag{26}$$

The contraction property of the $\mathcal{M}$ operator with respect to this norm can then be easily established (see (Bertsekas, 1995)): $\|\mathcal{M}\mathbf{q}_1 - \mathcal{M}\mathbf{q}_2\|_\infty \le \gamma\|\mathbf{q}_1 - \mathbf{q}_2\|_\infty$. 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 $\mathcal{M}\mathbf{q}_* = \mathbf{q}_*$ (Bertsekas, 1995).

Dual Representation. To establish convergence of the off-policy update $\mathcal{M}$ in the dual representation, we first define a form of max-norm for state-action visit distributions by

$$\|H\|_{\infty,\mathbf{r}} \quad = \quad \max_{(sa)} \Big| \sum_{(s'a')} H_{(sa,s'a')}\mathbf{r}_{(s'a')} \Big| \tag{27}$$

Then convergence of $\mathcal{M}H$ can be established by reducing the dual to the primal case by appealing to their relationship.

Lemma 19 *If* $(1-\gamma)\mathbf{q} = H\mathbf{r}$*, then* $(1-\gamma)\mathcal{M}\mathbf{q} = \mathcal{M}H\mathbf{r}$.

Proof Given the assumption $(1-\gamma)\mathbf{q} = H\mathbf{r}$ it follows that $(1-\gamma)\mathbf{q}_{(sa)} = [H\mathbf{r}]_{(sa)}$ for all sa. Then together with the definitions (21) and (22), we obtain $(1-\gamma)\mathcal{M}\mathbf{q} = (1-\gamma)(\mathbf{r} + \gamma P\Pi^*[\mathbf{q}]) = (1-\gamma)\mathbf{r} + \gamma P\Pi^*[(1-\gamma)\mathbf{q}] = (1-\gamma)\mathbf{r} + \gamma P\Pi^*[H\mathbf{r}] = (1-\gamma)\mathbf{r} + \gamma P\Pi^*_\mathbf{r}[H]\mathbf{r} = ((1-\gamma)I + \gamma P\Pi^*_\mathbf{r}[H])\,\mathbf{r} = \mathcal{M}H\mathbf{r}$. ■

Thus, each $\mathcal{M}$ update preserves the relationship between H and $\mathbf{q}$. Therefore, given convergence of $\mathcal{M}\mathbf{q}$ to a fixed point $\mathbf{q}_*$, $\mathcal{M}\mathbf{q}_* = \mathbf{q}_*$, it must hold that $\mathcal{M}H$ converges to

H_* such that $H_*\mathbf{r} = (1-\gamma)\mathbf{q}_*$. Again, one subtlety here is that the dual fixed point is not unique. This is, once again, not a contradiction because the norm on dual representations $\|\cdot\|_{\infty,\mathbf{r}}$ 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. Thus the off-policy operator converges to a simplex of equivalent fixed points $\{H_* : \mathcal{M}H_* = H_*\}$.

6. Approximate Dynamic Programming

Scaling up DP and RL algorithms to large problem domains is one of the central challenges in RL research. The most common approach is to generalize state or state-action value functions using function approximation, where a target is approximated by a weighted combination of basis functions. Although the combination could be non-linear, it is most common to focus on linear approximations, which we also do here. Although primal and dual updates exhibit strong equivalence in the tabular case, important differences begin to emerge when one considers approximation.

6.1 Linear Approximation Schemes

For succinctness, we focus on the state-action case only.

Primal Representation. In the primal representation, linear function approximation proceeds, conceptually, by fixing a small set of k basis vectors in a $|S||A| \times k$ matrix Φ. Approximate state-action value vectors $\hat{\mathbf{q}}$ can then be expressed by linear combinations of these bases, $\hat{\mathbf{q}} = \Phi\mathbf{w}$, where $\mathbf{w}$ is a $k \times 1$ vector of adjustable weights. Thus, $\hat{\mathbf{q}} \in \text{span}(\Phi)$ (where it is understood that this refers to the column span).

Dual Representation. In the dual representation, a similar linear approximation strategy can be followed: one begins with a set of bases that are linearly combined to form an approximation $\hat{H}$. However, the resulting approximation $\hat{H}$ must satisfy the constraints that it is nonnegative and row normalized (cf. Lemma 10). Therefore, let $\boldsymbol{\Upsilon} = (\Upsilon^{(1)}, ..., \Upsilon^{(k)})$ denote a set of k basis matrices such that each $\Upsilon^{(i)}$ is an $|S||A| \times |S||A|$ matrix satisfying the constraints $\Upsilon^{(i)} \geq 0$ and $\Upsilon^{(i)}\mathbf{1} = \mathbf{1}$. Then a valid representation of a dual approximation $\hat{H}$ can be expressed by a convex combination

$$\hat{H} \quad = \quad w_1\Upsilon^{(1)} + \cdots + w_k\Upsilon^{(k)} \quad \text{subject to} \quad \mathbf{w} \geq 0,\, \mathbf{w}^\top\mathbf{1} = 1 \tag{28}$$

where $\mathbf{w}$ is once again a $k \times 1$ vector of adjustable weights. That is, $\hat{H} \in \text{simplex}(\boldsymbol{\Upsilon})$. It is easy to verify that this construction results in appropriately row normalized matrices.

Lemma 20 $\hat{H} \geq 0$ *and* $\hat{H}\mathbf{1} = \mathbf{1}$

Proof Nonnegativity is immediate since $\hat{H}$ is a convex combination of nonnegative matrices. Second, since $\mathbf{w}$ and each row of $\Upsilon^{(i)}$ are normalized, it follows that $\hat{H}\mathbf{1} = w_1\Upsilon^{(1)}\mathbf{1} + \cdots + w_k\Upsilon^{(k)}\mathbf{1} = w_1\mathbf{1} + \cdots + w_k\mathbf{1} = (\mathbf{w}^\top\mathbf{1})\mathbf{1} = \mathbf{1}$. ■

Below we will often find it convenient to work with a vectorized form of $\hat{H}$. That is, let Ψ be an $(|S||A|)^2 \times k$ matrix of basis vectors such that $\Psi_{(:,i)} = \mathrm{vec}(\Upsilon^{(i)})$. Then we can represent $\hat{H}$ in a vectorized form as

$$\hat{\mathbf{h}} \;=\; \mathrm{vec}(\hat{H}) \;=\; \Psi\mathbf{w} \;=\; \mathrm{vec}\Big(w_1\Upsilon^{(1)} + \cdots + w_k\Upsilon^{(k)}\Big) \tag{29}$$

To recover a matrix from a vector representation, one can use an inverse operator such that $\hat{H} = \mathrm{reshape}(\hat{\mathbf{h}}) = \mathrm{reshape}(\mathrm{vec}(\hat{H}))$. A valid vector basis Ψ can then be specified by any $(|S||A|)^2 \times k$ matrix such that $\Psi \geq 0$ and $(\mathbf{1}^\top \otimes I)\Psi = \mathbf{1}\mathbf{1}^\top$, where $\otimes$ denotes Kronecker product and $\mathbf{1}\mathbf{1}^\top$ is the matrix of all 1s.

Lemma 21 $\mathrm{vec}(\hat{H}) = \Psi\mathbf{w}$ *implies* $\hat{H} \geq 0$ *and* $\hat{H}\mathbf{1} = \mathbf{1}$

Proof Nonnegativity is obvious since $\Psi \geq 0$ and $\mathbf{w} \geq 0$ by assumption. Second, given the assumptions $(\mathbf{1}^\top \otimes I)\Psi = \mathbf{1}\mathbf{1}^\top$ and $\mathbf{w}^\top\mathbf{1} = 1$, it follows that $\hat{H}\mathbf{1} = \mathrm{vec}(\hat{H}\mathbf{1}) = (\mathbf{1}^\top \otimes I)\mathrm{vec}(\hat{H}) = (\mathbf{1}^\top \otimes I)\Psi\mathbf{w} = \mathbf{1}\mathbf{1}^\top\mathbf{w} = \mathbf{1}$. ■

Thus, one can work equivalently in the simplex of basis matrices, $\hat{H} \in \mathrm{simplex}(\boldsymbol{\Upsilon})$, or the corresponding simplex of basis vectors, $\hat{\mathbf{h}} \in \mathrm{simplex}(\Psi)$, depending on which form is most convenient.

6.2 Projection Operator

Recall that in the primal, the state-action vector $\mathbf{q}$ 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. Instead, a representable approximation is required. The subtlety resolved by Tsitsiklis and Van Roy (1997) is to identify a particular form of best approximation—weighted least squares with respect to the stationary distribution $\mathbf{z}$ (23)—that ensures convergence to a fixed point $\mathbf{q}_+$ is still achieved when approximation is combined with the on-policy operator $\mathcal{O}$. Unfortunately, there are a few shortcomings associated with using least squares projection. First, the fixed point $\mathbf{q}_+$ of the combined operator—$\mathcal{O}$ composed with projection—is not guaranteed to be the best representable approximation of $\mathcal{O}$'s fixed point, $\mathbf{q}_\Pi$. Instead, only a bound can be proven on how close this altered fixed point is to the best representable approximation. Second, it is well known that the off-policy update $\mathcal{M}$ does not always have a fixed point when combined with least squares projection in the primal (de Farias and Van Roy, 2000), and consequently suffers the risk of divergence (Baird, 1995; Sutton and Barto, 1998). A key advantage of the dual approach is that linear approximation cannot diverge, even with off-policy updates, due to boundedness. Third, exact projections do not permit practical algorithms because they require expectations to be computed over the entire state or state-action spaces. Nevertheless, projection provides a useful basis for analysis, and a foundation for deriving subsequent practical algorithms.

We first focus on establishing the main convergence results for projection with the on-policy operator $\mathcal{O}$.

Primal Representation. The main steps taken in proving a bound in the primal case are useful to proving a bound in the dual. First, a map from a general $\mathbf{q}$ vector onto its best approximation in span(Φ) can be defined by an operator $\mathcal{P}$ that projects $\mathbf{q}$ into the column span of Φ

$$\mathcal{P}\mathbf{q} \quad = \quad \underset{\hat{\mathbf{q}}\in\mathrm{span}(\Phi)}{\mathrm{argmin}} \; \|\mathbf{q}-\hat{\mathbf{q}}\|_{\mathbf{z}}{}^2 \quad = \quad \Phi(\Phi^\top Z\Phi)^{-1}\Phi^\top Z\mathbf{q} \tag{30}$$

where $Z = \mathrm{diag}(\mathbf{z})$. The crucial property of this weighted projection is that it is a non-expansion in $\|\cdot\|_{\mathbf{z}}$; that is, $\|\mathcal{P}\mathbf{q}\|_{\mathbf{z}} \le \|\mathbf{q}\|_{\mathbf{z}}$, which can be easily established using a generalized Pythagorean theorem. Approximate dynamic programming then proceeds by composing the two operators—the on-policy update $\mathcal{O}$ with the subspace projection $\mathcal{P}$—to compute the best representable approximation of the one step update. This combined operator is guaranteed to converge to a fixed point $\mathbf{q}_+$, since composing a non-expansion with a contraction is still a contraction. Tsitsiklis and Van Roy (1997) then use these facts to establish an approximation bound between $\mathbf{q}_+$ and the fixed point of the on-policy operator, $\mathbf{q}_\Pi$: $\|\mathbf{q}_+ - \mathbf{q}_\Pi\|_{\mathbf{z}} \le \frac{1}{1-\gamma}\|\mathbf{q}_\Pi - \mathcal{P}\mathbf{q}_\Pi\|_{\mathbf{z}}$.

Dual Representation. The dual case is somewhat more complicated because one needs to represent nonnegative, row normalized matrices, not just vectors. Nevertheless, a very similar approach to the primal case can be applied successfully. In the dual, the state-action visit distribution H is approximated by a linear combination of basis matrices in $\boldsymbol{\Upsilon}$. Once again, there is no reason to expect an update like $\mathcal{O}H$ or $\mathcal{M}H$ to keep the matrix in simplex($\boldsymbol{\Upsilon}$). Therefore, a projection operator must be constructed that determines a best representable approximation. This projection needs to be defined with respect to the right norm to ensure convergence, which can be achieved by the pseudo-norm $\|\cdot\|_{\mathbf{z},\mathbf{r}}$ defined in (25). Accordingly, define the weighted projection operator $\mathcal{P}$ over matrices by

$$\mathcal{P}H \quad = \quad \underset{\hat{H}\in\mathrm{simplex}(\boldsymbol{\Upsilon})}{\mathrm{argmin}} \quad \|H-\hat{H}\|_{\mathbf{z},\mathbf{r}}{}^2 \tag{31}$$

This projection can be computed by solving a quadratic program

$$\hat{\mathbf{w}} \quad = \quad \underset{\mathbf{w}\ge 0,\, \mathbf{w}^\top\mathbf{1}=1}{\mathrm{argmin}} \; \|H - \mathrm{reshape}(\Psi\mathbf{w})\|_{\mathbf{z},\mathbf{r}}{}^2$$

and recovering $\hat{H} = \mathrm{reshape}(\Psi\hat{\mathbf{w}})$. A key result is that this projection operator is a non-expansion with respect to the pseudo-norm $\|\cdot\|_{\mathbf{z},\mathbf{r}}$.

Theorem 22 $\|\mathcal{P}H\|_{\mathbf{z},\mathbf{r}} \;\le\; \|H\|_{\mathbf{z},\mathbf{r}}$

Proof Note that the projection operator $\mathcal{P}$ can be viewed as a composition of three orthogonal projections: first, onto the linear subspace span($\boldsymbol{\Upsilon}$), then onto the subspace of row normalized matrices span($\boldsymbol{\Upsilon}$)$\cap\{H : H\mathbf{1} = \mathbf{1}\}$, and finally onto the space of nonnegative matrices span($\boldsymbol{\Upsilon}$) $\cap \{H : H\mathbf{1} = \mathbf{1}\} \cap \{H : H \ge 0\}$. Note that the last projection into the nonnegative half-space 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: they satisfy the following generalized Pythagorean theorem. Consider just one of these linear projections $\mathcal{P}_1$

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

hence $\|\mathcal{P}_1 H\|_{\mathbf{z},\mathbf{r}} \leq \|H\|_{\mathbf{z},\mathbf{r}}$. Since the overall projection $\mathcal{P} = \mathcal{P}_1 \circ \mathcal{P}_2 \circ \mathcal{P}_3$ 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 result analogous to the primal case, which bounds the approximation error between H_+ and the best representable approximation to the on-policy fixed point H_Π, where $H_\Pi = \mathcal{O}H_\Pi$.

Theorem 23 $\|H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}} \;\leq\; \frac{1}{1-\gamma}\|\mathcal{P}H_\Pi - H_\Pi\|_{\mathbf{z},\mathbf{r}}$

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 the generalized Pythagorean theorem. Then since $H_+ = \mathcal{PO}H_+$ and $\mathcal{P}$ is a non-expansion, we have $\|H_+ - \mathcal{P}H_\Pi\|_{\mathbf{z},\mathbf{r}} = \|\mathcal{PO}H_+ - \mathcal{P}H_\Pi\|_{\mathbf{z},\mathbf{r}} \leq \|\mathcal{O}H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}}$. Finally, using the fact that $H_\Pi = \mathcal{O}H_\Pi$, Lemma 18 can be used to establish $\|\mathcal{O}H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}} = \|\mathcal{O}H_+ - \mathcal{O}H_\Pi\|_{\mathbf{z},\mathbf{r}} \leq \gamma\|H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}}$. Therefore $(1-\gamma)\|H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}} \leq \|\mathcal{P}H_\Pi - H_\Pi\|_{\mathbf{z},\mathbf{r}}$, which gives the result. ■

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 norm. Automatically, the dual updates cannot diverge, even under the composition $\mathcal{PM}$.

6.3 Gradient Operator

For large scale problems one does not normally have the luxury of computing full parallel DP updates. Furthermore, least squares projections also require knowing the stationary distribution $\mathbf{z}$ for $P\Pi$ (essentially requiring one to know the model of the MDP). A key intermediate step toward achieving practical algorithms is to formulate a gradient step operator that only approximates full projections. Unfortunately, practicality comes with a cost. As we will see in Section 7 below, a gradient step operator causes significant instability when composed with the off-policy update in the primal representation, to the extent that divergence is a common phenomenon. Fortunately, divergence is not possible in the dual representation; moreover all DP algorithms appear to converge reliably in the dual case.

Primal Representation. Gradient step updates are easily derived from a given projection operator. In the primal representation, projection is equivalent to solving for a weight vector $\mathbf{w}$ that minimizes the least squares objective $J = \frac{1}{2}\|\mathbf{q} - \hat{\mathbf{q}}\|_{\mathbf{z}}{}^2 = \frac{1}{2}\|\mathbf{q} - \Phi\mathbf{w}\|_{\mathbf{z}}{}^2$. Using the relation $\hat{\mathbf{q}} = \Phi\mathbf{w}$, we can derive the gradient update as $\mathcal{G}_{\hat{\mathbf{q}}}\mathbf{q} = \hat{\mathbf{q}} - \alpha\Phi\nabla_{\mathbf{w}}J = \hat{\mathbf{q}} - \alpha\Phi\Phi^\top Z(\hat{\mathbf{q}} - \mathbf{q})$, where α is a positive step-size parameter. Here, the target vector $\mathbf{q}$ is usually given by a DP update (either $\mathcal{O}$ or $\mathcal{M}$) to a representable vector $\hat{\mathbf{q}}$. This gives the composed updates

$$\mathcal{G}\mathcal{O}\hat{\mathbf{q}} \;=\; \hat{\mathbf{q}} - \alpha\Phi\Phi^\top Z(\hat{\mathbf{q}} - \mathcal{O}\hat{\mathbf{q}}) \tag{32}$$

$$\mathcal{G}\mathcal{M}\hat{\mathbf{q}} \;=\; \hat{\mathbf{q}} - \alpha\Phi\Phi^\top Z(\hat{\mathbf{q}} - \mathcal{M}\hat{\mathbf{q}}) \tag{33}$$

for the on-policy and off-policy cases respectively. In fact, these are parallel versions of the standard RL updates with function approximation (see Section 6.4). Our experimental results in Section 7 show that the gradient update is stable when composed with the on-policy operator (32), but usually diverges when composed with the off-policy update (33).

Dual Representation. In the dual representation, one can derive a gradient update similarly, except that simplex constraints must be maintained on $\mathbf{w}$. Consider the objective

$$J \;=\; \frac{1}{2}\|H - \hat{H}\|_{\mathbf{z},\mathbf{r}}{}^2 \;=\; \frac{1}{2}\|\mathrm{vec}(H\mathbf{r}) - \mathrm{vec}(\hat{H}\mathbf{r})\|_{\mathbf{z}}{}^2 \;=\; \frac{1}{2}\|(\mathbf{r}^\top \otimes I)(\mathbf{h} - \Psi\mathbf{w})\|_{\mathbf{z}}{}^2 \tag{34}$$

The unconstrained gradient with respect to $\mathbf{w}$ is given by

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

where $\Gamma = (\mathbf{r}^\top \otimes I)\Psi$. This gradient direction cannot be followed directly because it might violate the simplex constraints. Fortunately, enforcing these constraints is easy. For example, the constraint $\mathbf{w}^\top\mathbf{1} = \mathbf{1}$ can be maintained by first projecting the gradient onto the constraint, obtaining the modified update direction $\delta\mathbf{w} = (I - \frac{1}{k}\mathbf{1}\mathbf{1}^\top)\nabla_{\mathbf{w}}J$. In this case, the weight vector can be updated by

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

where α is a positive step-size parameter. Subsequently, if the update violates any nonnegativity constraint, the step-size can be shortened and the update direction can further be projected onto the simplex boundaries (we omit these details for succinctness of presentation). The gradient operator can then be defined by

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

Here, the target vector $\mathbf{h}$ is determined by the underlying DP update (either $\mathcal{O}$ or $\mathcal{M}$) to a representable vector $\hat{\mathbf{h}}$. This gives the composed updates

$$\mathcal{G}\mathcal{O}\hat{\mathbf{h}} \;=\; \hat{\mathbf{h}} - \alpha\Psi\Big(I - \frac{1}{k}\mathbf{1}\mathbf{1}^\top\Big)\Gamma^\top Z(\mathbf{r}^\top \otimes I)(\hat{\mathbf{h}} - \mathcal{O}\hat{\mathbf{h}}) \tag{35}$$

$$\mathcal{G}\mathcal{M}\hat{\mathbf{h}} \;=\; \hat{\mathbf{h}} - \alpha\Psi\Big(I - \frac{1}{k}\mathbf{1}\mathbf{1}^\top\Big)\Gamma^\top Z(\mathbf{r}^\top \otimes I)(\hat{\mathbf{h}} - \mathcal{M}\hat{\mathbf{h}}) \tag{36}$$

respectively for the on-policy and off-policy cases. We investigate the convergence properties of these composed updates experimentally in Section 7, and observe that they *both* converge reliably in the dual (whereas $\mathcal{GM}$ usually diverges in the primal).

6.4 Stochastic Gradient Operator

Finally, truly practical algorithms for large scale problems can be achieved by approximating the gradient operator with local stochastic estimates. That is, in place of a full expectation, the gradient updates can be estimated with a single sampled transition, $sa \to s'a'$. Such updates are *stochastic* gradient operators.

Primal Representation. The key observation in deriving an *unbiased* stochastic gradient update is to note that the projection gradient can be written as an expectation: $\nabla_{\mathbf{w}} J = \Phi^\top Z(\hat{\mathbf{q}} - \mathbf{q}) = \mathrm{E}_{(sa)\sim Z}\left[\Phi_{(sa,:)}^\top(\hat{\mathbf{q}}_{(sa)} - \mathbf{q}_{(sa)})\right]$. In an RL context, if states are visited according to a stationary exploration policy, Π, then each step $sa \to s'a'$ provides an unbiased sample of the gradient: $\widetilde{\nabla_{\mathbf{w}} J} = \Phi_{(sa,:)}^\top(\hat{\mathbf{q}}_{(sa)} - \mathbf{q}_{(sa)})$. Replacing expectations with unbiased samples and working strictly with the weight vector $\mathbf{w}$ yields the updates

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha\Phi_{(sa,:)}^\top(\Phi_{(sa,:)}\mathbf{w}_t - \mathbf{r}_{(sa)} - \gamma\Phi_{(s'a',:)}\mathbf{w}_t) \quad (37)$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha\Phi_{(sa,:)}^\top(\Phi_{(sa,:)}\mathbf{w}_t - \mathbf{r}_{(sa)} - \gamma\max_{a^*}\Phi_{(s'a^*,:)}\mathbf{w}_t) \quad (38)$$

for the on-policy and off-policy cases respectively. The former is the standard RL update for policy evaluation in large domains; the latter is the well known but unstable update, Q-learning with linear function approximation (Sutton and Barto, 1998).

Dual Representation. In the dual, the gradient of the objective can also be written as an expectation

$$\nabla_{\mathbf{w}} J = \Gamma^\top Z(\mathbf{r}^\top \otimes I)(\hat{h} - h) = \mathrm{E}_{(sa)\sim Z}\left[\Gamma_{(sa,:)}^\top(\Gamma_{(sa,:)}\mathbf{w} - H_{(sa,:)}\mathbf{r})\right] \quad (39)$$

where $\Gamma = (\mathbf{r}^\top \otimes I)\Psi$. Then, once again, each step $sa \to s'a'$ taken by a stationary exploration policy Π provides an unbiased sample of the gradient: $\widetilde{\nabla_{\mathbf{w}} J} = \Gamma_{(sa,:)}^\top(\Gamma_{(sa,:)}\mathbf{w} - H_{(sa,:)}\mathbf{r})$. Replacing expectations with unbiased samples and working strictly with the weight vector $\mathbf{w}$ yields the dual versions of the RL updates

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha\Gamma_{(sa,:)}^\top\left(\Gamma_{(sa,:)}\mathbf{w}_t - (1-\gamma)\mathbf{r}_{(sa)} - \gamma\Gamma_{(s'a',:)}\mathbf{w}_t\right) \quad (40)$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \alpha\Gamma_{(sa,:)}^\top\left(\Gamma_{(sa,:)}\mathbf{w}_t - (1-\gamma)\mathbf{r}_{(sa)} - \max_{a^*}\gamma\Gamma_{(s'a^*,:)}\mathbf{w}_t\right) \quad (41)$$

for the on-policy case and off-policy cases respectively. (As before, these gradient updates need to be projected into the constraint simplex, which is not difficult, but we omit details for brevity.) The computational cost of the dual updates is not significantly greater than the primal. These updates provide practical dual algorithms for large scale problems. The major advantage of the dual updates versus the primal is that they cannot diverge. In particular, (41) is well behaved in practice, whereas (38) is highly unstable and tends to diverge, unless specific care is taken (Gordon, 1995; Sutton, 1996). We investigate the convergence properties of the various updates empirically below.

7. Experimental Results

To investigate the effectiveness of the dual representations, we conducted experiments with the various DP algorithms on different problem domains. In particular, we considered three distinct problem domains: randomly synthesized MDPs, Baird's *star problem*, and the *mountain car* problem. The randomly synthesized MDP problems allow us to test the general properties of the algorithms. Baird's star problem is perhaps the most-cited example of a problem where Q-learning with linear function approximation diverges (Baird, 1995). The mountain car domain has been prone to divergence with some primal representations (Boyan and Moore, 1995), although successful results have been reported by carefully choosing bases based on sparse tile coding (Sutton, 1996).

For each problem domain, we experimented with twelve dynamic programming algorithms: tabular on-policy ($\mathcal{O}$), projected on-policy ($\mathcal{PO}$), gradient on-policy ($\mathcal{GO}$), tabular off-policy ($\mathcal{M}$), projected off-policy ($\mathcal{PM}$), and gradient off-policy ($\mathcal{GM}$), for both the primal and dual representations. (We did not include the stochastic gradient updates developed in Section 6.4, since their behavior is expected to follow that of the gradient operators, due to the fact that they are unbiased estimates.) In each case, the algorithms were run with 100 repeats to a horizon of 1000 iterations. The discount factor was set to $\gamma = 0.9$. The step size for the gradient updates was 0.1 for primal representations and 100 for dual representations. Unless otherwise specified, the initial values of state-action value functions $\mathbf{q}$ are set according to standard normal distribution and state-action visit distributions H are chosen uniformly randomly with row normalization.

The plots show error obtained versus the number of DP iterations executed. For the on-policy algorithms, the errors are measured between the current estimates (either $\mathbf{q}$ or H) and the optimal fixed point determined by the policy, using the norm defined in (24) and (25) for the primal and dual cases, respectively. For the off-policy algorithms, the errors are measured between the current estimates (either $\mathbf{q}$ or H) and the optimal solutions (either $\mathbf{q}^*$ or H^*), using the max norm defined in (26) and (27) for the primal and dual cases, respectively. Figures 1, 4, and 6 show the behavior of the on-policy update operators on different problems, and Figures 2, 5, and 7 show the behavior of the off-policy update operators.

7.1 Task: Randomly Synthesized MDPs

For the synthetic MDPs, we generated the transition model P and reward function $\mathbf{r}$ randomly—the transition function is uniformly distributed between 0 and 1 and the reward function is normally distributed with mean 0 and variance 1. Since our goal is to test the stability of algorithms without carefully crafting features, we also choose random basis functions according to standard normal distribution for primal representations, and random basis distributions according to uniform distribution for dual representations. Although we conducted experiments on other problem sizes, we only report results for random MDPs with 100 states, 5 actions, and 10 bases, averaging over 100 repeats here. We observed consistent behavior of the algorithms over ensembles of random MDPs, across different number of states, actions, and bases.

Figure 2 shows that the gradient off-policy ($\mathcal{GM}$) algorithm diverges, while all the other algorithms in Figures 1 and 2 converge.

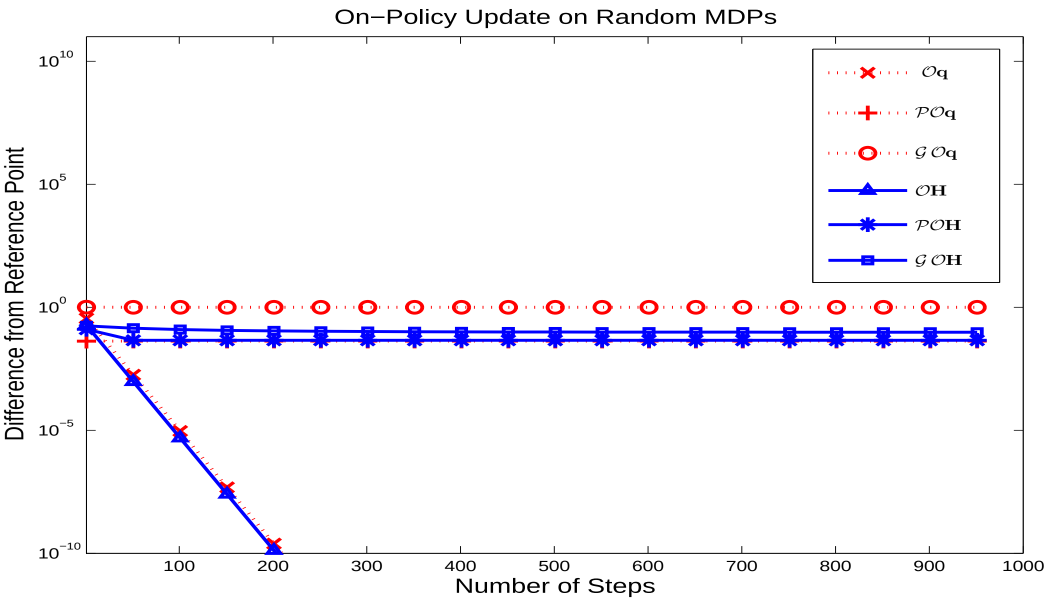

Figure 1: On-policy update of state-action value $\mathbf{q}$ and visit distribution H on randomly synthesized MDPs

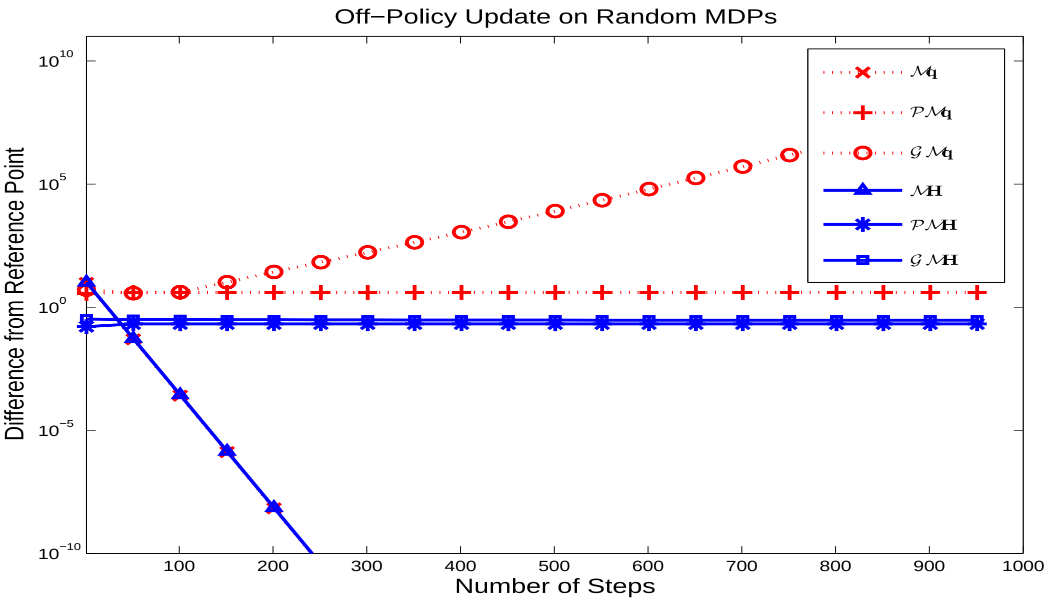

Figure 2: Off-policy update of state-action value $\mathbf{q}$ and visit distribution H on randomly synthesized MDPs

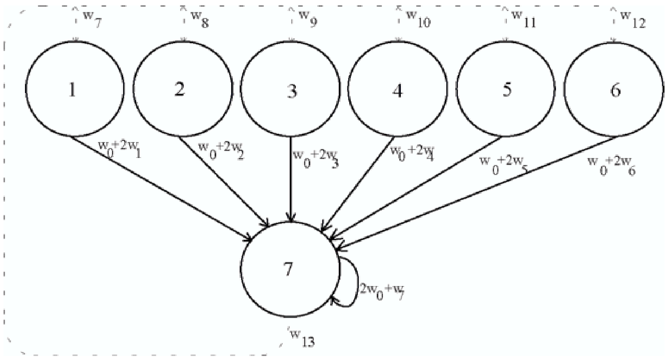

Figure 3: Baird's star problem (Baird, 1995)

7.2 Task: Baird's Star Problem

Baird's star problem, shown in Figure 3, has 7 states and 2 actions. The reward function is uniformly zero. The transitions in solid lines are triggered by action a_1 and the transitions in dotted lines are generated by action a_2, that is, taking action a_1 in all the states will cause a transition to state s_7 with probability 1; taking action a_2 in each state will cause a transition to one of states s_1 through s_6 with equal probability 1/6.

In our experiments, we used the same exploration policy and linear value function approximation as (Baird, 1995). The fixed policy chooses action a_1 with probability 1/7 and action a_2 with probability 6/7. The representation of action values for the primal case is given in Figure 3. The action values are given by the linear combination of the weights. For example, $Q(s_1, a_1) = w_0 + 2w_1$ and $Q(s_1, a_2) = w_7$; see Figure 3. The initial action values of a_1 are set to be bigger than the initial values of a_2, and the value of action a_1 in state six is set to the largest action value.

Note that for the dual approach, all of the updates obtain exactly zero error in this problem domain, according to the pseudo-norms defined in (25) and (27), since $\mathbf{r} = 0$. Therefore, we do not plot any dual results in this case. However, for the primal approach, the gradient off-policy update diverges, as shown by the dotted line with the circle marker in Figure 5. Convergence is obtained in the on-policy case.

7.3 Task: The Mountain Car Problem

The mountain car problem has continuous state and action spaces, which we discretize with a simple grid, resulting in an MDP with 222 states and 3 actions. The number of bases is chosen to be 5 in both the primal and dual representations. We chose the bases for the algorithms randomly. In the primal representation, we randomly generated basis functions according to a standard normal distribution. In the dual representation, we randomly generated basis distributions according to a uniform distribution, and renormalized.

Figure 7 again shows that the gradient off-policy update in the primal diverges, while all of the dual algorithms converge (see Figures 6 and 7).

Overall, we note that the approximate dual algorithms tended to achieve smaller approximation errors than the corresponding primal algorithms. Furthermore, in our experiments

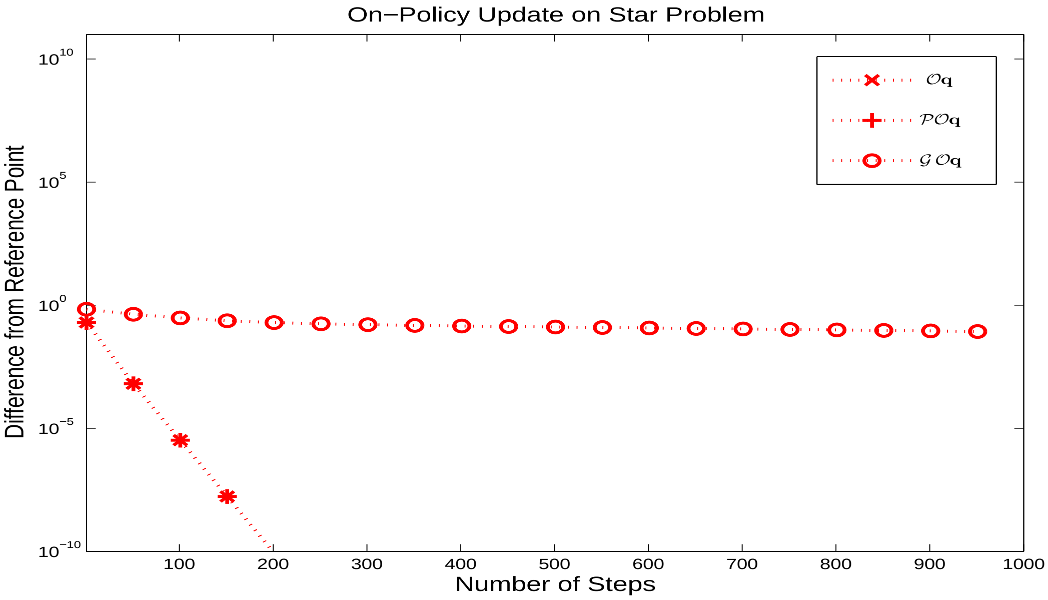

Figure 4: On-policy update of state-action value **q** on Baird's star problem

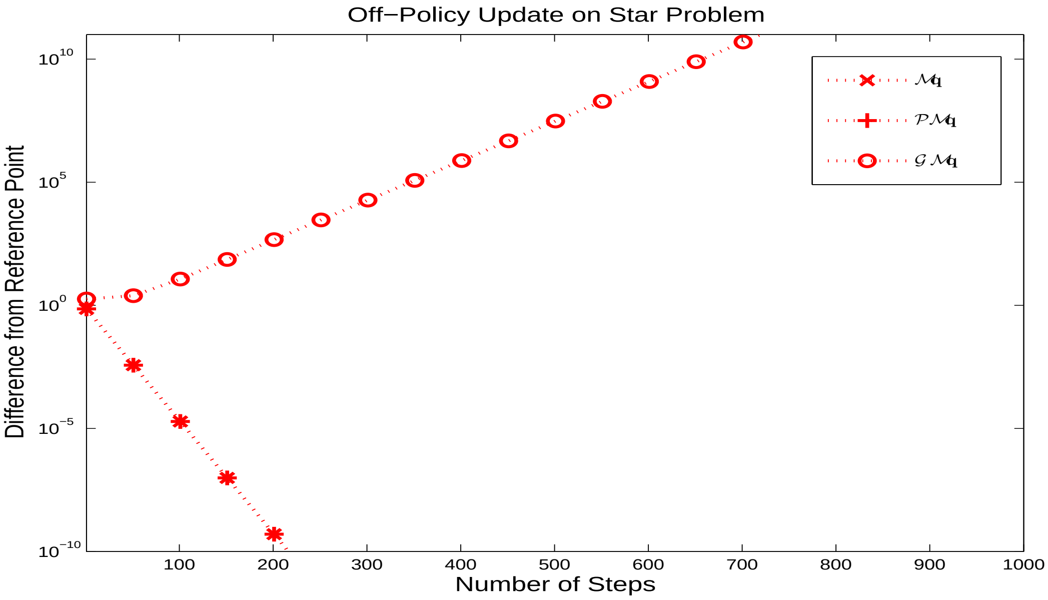

Figure 5: Off-policy update of state-action value **q** Baird's star problem

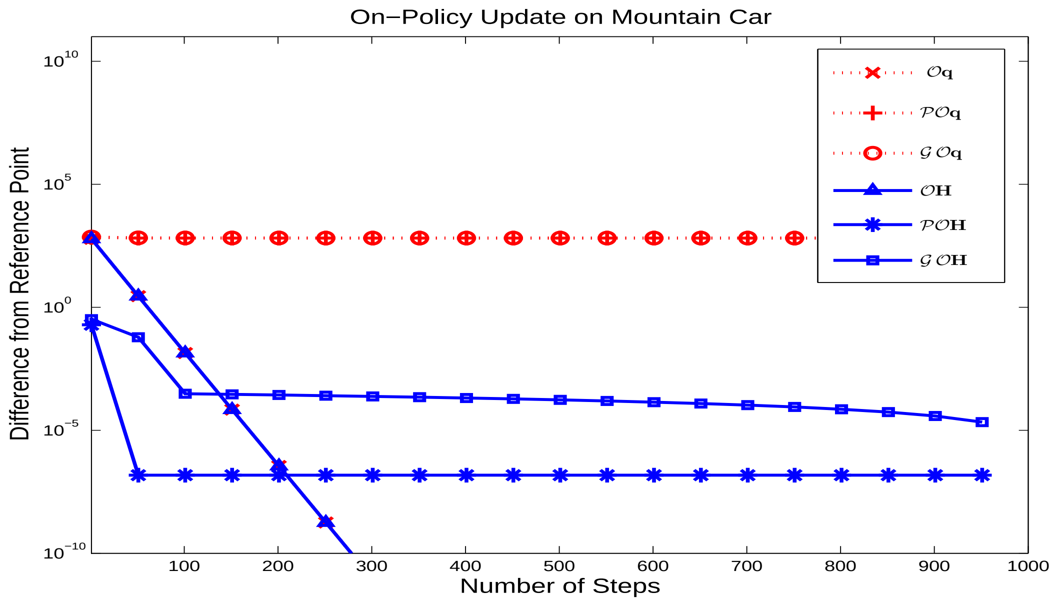

Figure 6: On-policy update of state-action value $\mathbf{q}$ and visit distribution H on the mountain car problem

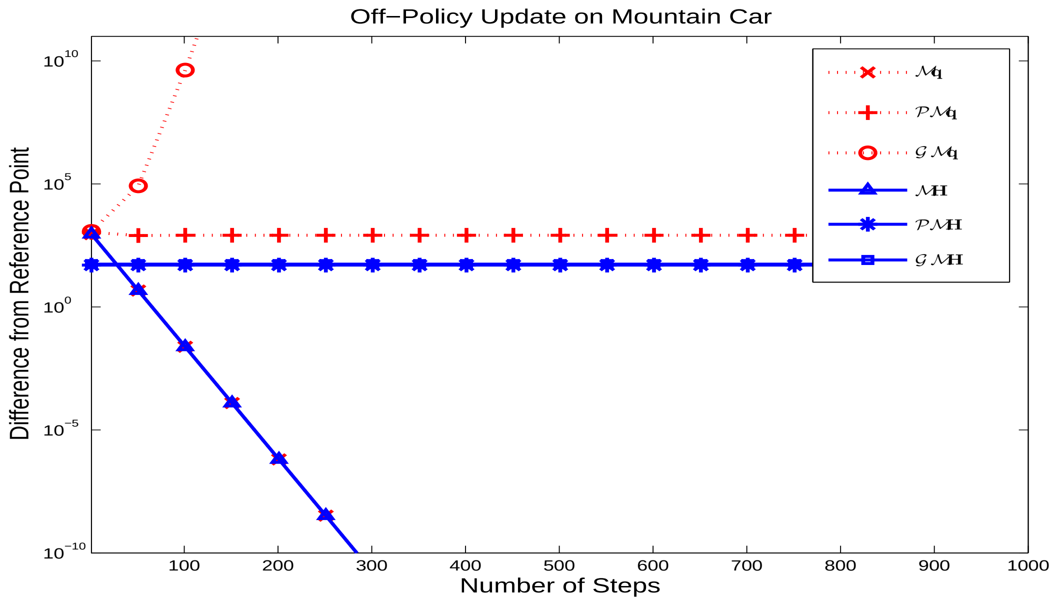

Figure 7: Off-policy update of state-action value $\mathbf{q}$ and visit distribution H on the mountain car problem

the dual DP algorithms always converged, while the primal DP algorithms were only stable as long as gradient updates were not composed with the off-policy operator. In particular, in the primal representation, the gradient off-policy update with linear function approximation almost always diverged.

8. Conclusion

We have introduced a series of dual DP algorithms for sequential decision making problems based on maintaining an explicit representation of visit distributions as opposed to value functions. We studied the convergence properties of these dual algorithms both theoretically and empirically, and found that the on-policy updates converge in both primal and dual algorithms, while the off-policy updates diverge when composed with gradient operator in the primal but converged in the dual. In particular, we proved the convergence of tabular on-policy ($\mathcal{O}$), tabular off-policy ($\mathcal{M}$), and projected on-policy ($\mathcal{PO}$) updates for the dual, while only establishing experimentally the convergence of the projected off-policy ($\mathcal{PM}$), gradient on-policy ($\mathcal{GO}$), and gradient off-policy ($\mathcal{GM}$) in the dual.

A potential limitation of the dual approach is that, in the tabular case, the updates in the dual representation are more expensive than in the primal representation. However, this limitation can be largely overcome by exploiting function approximation with stochastic gradient updates (Section 6.4). We plan to investigate the convergence properties of these practical RL algorithms both theoretically and empirically in future work.

Overall the dual approach offers a coherent and comprehensive perspective on optimal sequential decision making problems, and provides a viable alternative to standard value function based approaches for developing DP and RL algorithms. An interesting opportunity we have not yet explored is to investigate joint primal-dual algorithms that might permit tighter bounds on approximation quality to be obtained.

Acknowledgments

This work was performed while the first author was in the Department of Computing Science, University of Alberta. We gratefully acknowledge the support of Alberta Ingenuity, CRC, NSERC, MITACS and CFI.

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.

Geoffrey J. Gordon. Stable function approximation in dynamic programming. In *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*, 1995.

Michail G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003. ISSN 1533-7928.

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. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in Neural Information Processing Systems*, pages 1038–1044, 1996.

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.

Tao Wang. *New Representations and Approximations for Sequential Decision Making under Uncertainty.* PhD thesis, University of Alberta, 2007.

Tao Wang, Michael Bowling, and Dale Schuurmans. Dual representations for dynamic programming and reinforcement learning. In *Proceedings of the 2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 44–51, 2007.

Tao Wang, Daniel Lizotte, Michael Bowling, and Dale Schuurmans. Stable dual dynamic programming. In *Proceedings of Advances in Neural Information Processing Systems 20*, 2008.