Stable Dynamic Programming and Reinforcement Learning with Dual Representations

Dynamic programming, reinforcement learning, approximation

Abstract

We investigate novel, dual algorithms for dynamic programming and reinforcement learning, based on maintaining explicit representations of stationary distributions instead of value functions. In particular, we investigate the convergence properties of standard dynamic programming and reinforcement learning algorithms when they are converted to their natural dual form. Here we uncover advantages for the dual approach: dual update algorithms, since they are based on estimating normalized probability distributions rather than unbounded value functions, avoid divergence even in the presence of function approximation and off-policy updates. Moreover, dual update algorithms remain stable in situations where standard value function estimation diverges.

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 & 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 & 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. Just such an approach has been recently proposed in (Wang et al., 2007), although no analysis of convergence properties nor implementation of the ideas were investigated.

In this paper, we investigate the convergence properties of these newly proposed dual solution techniques that are based on representing state visit and state-action visit distributions instead of value functions. Here 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. The dual approach appears to hold a significant advantage over the standard primal view of DP/RL in one major sense: 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 in fact converge (i.e. avoid oscillation) in the very circumstance where primal updates can and often do diverge: gradient-based off-policy updates with linear function approximation (Baird, 1995; Sutton & Barto, 1998).

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 as a set of actions A, a set of states S, a $|S||A|$ by $|S|$ transition matrix P, a reward vector $\mathbf{r}$ and a discount factor γ. We address the discounted reward MDP formulation where the optimality criterion is maximizing the infinite horizon *discounted* reward $r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots = \sum_{t=0}^{\infty} \gamma^t r_t$. It is known that an optimal behavior strategy can always be expressed by a stationary *policy*, which we represent as an $|S||A| \times 1$ *vector* $\boldsymbol{\pi}$, whose entries $\boldsymbol{\pi}_{(sa)}$ specify the probability of taking action a in state s.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

The main problem is to compute an optimal policy given either (1) a complete specification of the environmental variables P and $\mathbf{r}$ (the "*planning problem*"), or (2) 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 first problem is normally tackled by LP or DP methods, and the second by RL 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 $\boldsymbol{\mu}$.

A standard LP for solving the planning problem can be expressed as

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

Here, Ξ is the $|S| \times |S||A|$ marginalization matrix; it is a sparse matrix built by placing $|S|$ row blocks of length $|A|$ in a block diagonal fashion, where each row block consists of all 1s. It is known that the optimal solution $\mathbf{v}^*$ to this LP corresponds to the *value function* for the optimal policy (Bertsekas, 1995; Bertsekas & Tsitsiklis, 1996). In particular, given $\mathbf{v}^*$, the optimal policy can be recovered by (5).

The dual LP can be derived by using Lagrange multipliers $\mathbf{d}$, a $|S||A| \times 1$ vector

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

Interestingly, we proved that any feasible vector in (2) is guaranteed to be normalized (Wang et al., 2007), and therefore the solution $\mathbf{d}^*$ is always a joint *probability distribution* over state-action pairs.

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 $\boldsymbol{\pi}^*_{(sa)} = \mathbf{d}^*_{(sa)} / \sum_a \mathbf{d}^*_{(sa)}$ (Ross, 1997).

4. Dual Representations

Dynamic programming methods for solving MDP evaluation and planning problems are typically expressed in terms of the primal value function. We demonstrated that all of these classical algorithms have natural duals expressed in terms of state and state-action probability distributions (Wang et al., 2007). Here we will only highlight key observations for the analysis of convergence in the following sections.

Policy Evaluation 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 or its distribution of discounted state visits. Below we will find it convenient to re-express a policy $\boldsymbol{\pi}$ by an equivalent representation as an $|S| \times |S||A|$ matrix Π where

$$\Pi_{(s,s'a)} \;=\; \begin{cases} \boldsymbol{\pi}_{(sa)} & \text{if } s' = s \\ 0 & \text{if } s' \neq s \end{cases}$$

One can quickly verify that the matrix product ΠP gives the *state to state* transition probabilities induced by the policy $\boldsymbol{\pi}$ in the environment P, and that $P\Pi$ gives the *state-action to state-action* transition probabilities induced by policy $\boldsymbol{\pi}$ in P.

When we consider RL algorithms below we will generally need to maintain joint *state-action* based evaluations. In the primal representation, the policy state-action value function can be specified by an $|S||A| \times 1$ vector $\mathbf{q} = \sum_{i=0}^{\infty} \gamma^i (P\Pi)^i \mathbf{r}$ which satisfies $\mathbf{q} = \mathbf{r} + \gamma P \Pi \mathbf{q}$.

To develop a *dual* form of state-action policy evaluation, consider the linear system

$$\mathbf{d}^\top \;=\; (1-\gamma)\boldsymbol{\nu}^\top + \gamma \mathbf{d}^\top P \Pi \tag{3}$$

where $\boldsymbol{\nu}$ is the initial distribution over state-action pairs. Not only is $\mathbf{d}$ a proper probability distribution over state-action pairs, it also allows one to easily compute the expected discounted return of the policy $\boldsymbol{\pi}$. However, recovering the state-action distribution $\mathbf{d}$ is inadequate for *policy improvement*, We therefore consider the following $|S||A| \times |S||A|$ matrix

$$H \;=\; (1-\gamma) I + \gamma H P \Pi \tag{4}$$

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). Unlike $\mathbf{d}^\top$ however, H drops the dependence on $\boldsymbol{\mu}$, giving $(1-\gamma)\mathbf{q} = H\mathbf{r}$. That is, given H we can easily recover the state-action values of $\boldsymbol{\pi}$.

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.

Given a current policy $\boldsymbol{\pi}$, whose state-action value function $\mathbf{q}$ have already been determined, one can derive an improved policy $\boldsymbol{\pi}'$ via the update

$$a^*(s) = \arg\max_a \mathbf{q}_{(sa)} \quad (5)$$

$$\boldsymbol{\pi}'_{(sa)} = \begin{cases} 1 & \text{if } a = a^*(s) \\ 0 & \text{if } a \neq a^*(s) \end{cases} \quad (6)$$

One can verify that this update leads to an improved policy (Wang et al., 2007).

This development can be parallelled in the dual by first defining and analogous policy update. Given a policy $\boldsymbol{\pi}$, the dual form of the policy update can be expressed in terms of the state-action matrix H for $\boldsymbol{\pi}$

$$a^*(s) = \arg\max_a H_{(sa,:)}\mathbf{r} \quad (7)$$

$$\boldsymbol{\pi}'_{(sa)} = \begin{cases} 1 & \text{if } a = a^*(s) \\ 0 & \text{if } a \neq a^*(s) \end{cases} \quad (8)$$

In fact, since $(1-\gamma)\mathbf{q} = H\mathbf{r}$, the two policy updates given in (5) and (7) respectively, must lead to the same resulting policy $\boldsymbol{\pi}'$. Further details are given in (Wang et al., 2007).

5. DP algorithms and convergence

We first investigate whether dynamic programming operators with the dual representations exhibit the same (or better) convergence properties to their primal counterparts. These questions will be answered in the affirmative, largely showing equivalence to the standard primal cases. The real advantage of the dual approach will arise below when we consider function approximation. To keep the presentation efficient, we will concentrate only on state-action based representations, $\mathbf{q}$ and H, respectively.

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. We will focus on two standard operators, the on-policy update and the max-policy update.

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

$$\begin{aligned} \mathcal{O}\mathbf{q} &= \mathbf{r} + \gamma P\Pi\mathbf{q} \\ \mathcal{O}H &= (1-\gamma)I + \gamma P\Pi H \end{aligned}$$

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

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

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 $H = (1-\gamma)I + \gamma P\Pi^*_r[H]$.

5.1. 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. Although these results are already well known, we repeat some brief details that will be helpful later.

First, to establish contraction, 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 $P\Pi$. (Note that $\mathbf{z}$ is not the same as the initial distribution $\boldsymbol{\nu}$ nor the discounted stationary distribution $\mathbf{d}$.) Let $Z = \text{diag}(\mathbf{z})$. Then define the norm

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

Crucially, for this norm, a state-action transition is not an expansion.

Lemma 1 $\|P\Pi\mathbf{q}\|_{\mathbf{z}} \leq \|\mathbf{q}\|_{\mathbf{z}}$ *(Tsitsiklis & Van Roy, 1997)*

Proof: The result follows from Jensen's inequality

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

This allows one to easily recover the fact that $\mathcal{O}$ is in fact a contraction with respect to $\|\cdot\|_{\mathbf{z}}$ in the primal case.

Lemma 2 $\|\mathcal{O}\mathbf{q}_1 - \mathcal{O}\mathbf{q}_2\|_{\mathbf{z}} \leq \gamma\|\mathbf{q}_1 - \mathbf{q}_2\|_{\mathbf{z}}$ *(Tsitsiklis & Van Roy, 1997)*

Proof:

$$\begin{aligned}\|\mathcal{O}\mathbf{q}_1 - \mathcal{O}\mathbf{q}_2\|_{\mathbf{z}} &= \|\mathbf{r} + \gamma P\Pi\mathbf{q}_1 - \mathbf{r} - \gamma P\Pi\mathbf{q}_2\|_{\mathbf{z}} \\ &= \gamma\|P\Pi(\mathbf{q}_1 - \mathbf{q}_2)\|_{\mathbf{z}} \leq \gamma\|\mathbf{q}_1 - \mathbf{q}_2\|_{\mathbf{z}}\end{aligned}$$

by Lemma 1. ∎

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, $\mathbf{q}_\Pi$ satisfied the policy based Bellman equation.

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

$$\|H\|_{\mathbf{z},\mathbf{r}}{}^2 = \|H\mathbf{r}\|_{\mathbf{z}}{}^2$$

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 3 $\|P\Pi H\|_{\mathbf{z},\mathbf{r}} \leq \|H\|_{\mathbf{z},\mathbf{r}}$

Proof:

$$\begin{aligned}\|P\Pi H\|_{\mathbf{z},\mathbf{r}} &= \|P\Pi(H\mathbf{r})\|_{\mathbf{z}} \\ &\leq \|H\mathbf{r}\|_{\mathbf{z}} = \|H\|_{\mathbf{z},\mathbf{r}}\end{aligned}$$

by Lemma 1. ∎

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

Lemma 4 $\|\mathcal{O}H_1 - \mathcal{O}H_2\|_{\mathbf{z},\mathbf{r}} \leq \gamma\|H_1 - H_2\|_{\mathbf{z},\mathbf{r}}$

Proof:

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

by Lemma 3. ∎

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 $\mathcal{O}$ 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.

5.2. 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.

Lemma 5 $\|\mathcal{M}\mathbf{q}_1 - \mathcal{M}\mathbf{q}_2\|_\infty \leq \gamma\|\mathbf{q}_1 - \mathbf{q}_2\|_\infty$ *(Bertsekas, 1995)*

The proof of this result is straightforward, but omitted for space. (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 max operator in the dual representation, we will simply reduce the dual to the primal case. Recall that there is a many to one relationship between the dual and primal representations, given by $H\mathbf{r} = (1-\gamma)\mathbf{q}$. To prove convergence of $\mathcal{M}H$, we simply appeal to this relationship.

Lemma 6 *If* $(1-\gamma)\mathbf{q} = H\mathbf{r}$*, then* $(1-\gamma)\mathcal{M}\mathbf{q} = \mathcal{M}H\mathbf{r}$.

Proof:

$$\begin{aligned}(1-\gamma)\mathcal{M}\mathbf{q} &= (1-\gamma)(\mathbf{r} + \gamma P\Pi^*[\mathbf{q}]) \\ &= (1-\gamma)\mathbf{r} + \gamma P\Pi^*[H\mathbf{r}] = \mathcal{M}H\mathbf{r}\end{aligned}$$

where the second equality holds since $(1-\gamma)\mathbf{q}_{(sa)} = [H\mathbf{r}]_{(sa)}$ for all (sa) by assumption. ∎

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 $H_*\mathbf{r} = (1-\gamma)\mathbf{q}_*$.

6. 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 & Van Roy, 2000), and consequently suffers the risk of divergence (Baird, 1995; Sutton & Barto, 1998).

Note that the max operator cannot diverge in the dual case, even with basis approximation, by boundedness alone; 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 basis approximation proceeds by fixing a small set of bases, forming a $|S||A| \times k$ matrix Φ, where k is the number of basis features. One then maintains the constraint that $\mathbf{q} \in \text{col\_span}(\Phi)$; i.e. that $\mathbf{q}$ can be expressed 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 & 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 $\mathcal{O}$.

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

$$\begin{aligned}\mathcal{P}\mathbf{q} &= \operatorname*{argmin}_{\mathbf{q}' \in \text{col\_span}(\Phi)} \|\mathbf{q} - \mathbf{q}'\|_2 \\ &= Q\mathbf{q} \quad \text{for} \quad Q = \Phi(\Phi^\top Z \Phi)^{-1}\Phi^\top Z\end{aligned}$$

The important property of this weighted projection is that it is a non-expansion in $\|\cdot\|_{\mathbf{z}}$.

Lemma 7 $\|\mathbf{q}\|_{\mathbf{z}}^2 = \|\mathcal{P}\mathbf{q}\|_{\mathbf{z}}^2 + \|\mathbf{q} - \mathcal{P}\mathbf{q}\|_{\mathbf{z}}^2$ *(Tsitsiklis & Van Roy, 1997)*

Immediately from this generalized Phythagorean theorem, which is reasonably easy to show, one obtains the non-expansion property $\|\mathcal{P}\mathbf{q}\|_{\mathbf{z}} \leq \|\mathbf{q}\|_{\mathbf{z}}$.

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. In fact, Lemma 2 can be re-established for the composition $\mathcal{P}\mathcal{O}$. Thus, by the contraction map fixed point theorem (Bertsekas, 1995) a fixed point must exist and is unique. Let $\mathbf{q}_+ = \mathcal{P}\mathcal{O}\mathbf{q}_+$ be the fixed point of the combined operator. Unfortunately, the fixed point $\mathbf{q}_+$ is not guaranteed to be the best representable approximation of $\mathcal{O}$'s fixed point $\mathbf{q}_\Pi$. Nevertheless, a bound can be proved on how close the altered fixed point $\mathbf{q}_+$ is to the best representable approximation $\mathcal{P}\mathbf{q}_\Pi$ of $\mathcal{O}$'s fixed point.

Lemma 8 $\|\mathbf{q}_+ - \mathbf{q}_\Pi\|_{\mathbf{z}} \leq \frac{1}{1-\gamma}\|\mathbf{q}_\Pi - \mathcal{P}\mathbf{q}_\Pi\|_{\mathbf{z}}$ *(Tsitsiklis & Van Roy, 1997)*

Proof: First note that $\|\mathbf{q}_+ - \mathbf{q}_\Pi\|_{\mathbf{z}} = \|\mathbf{q}_+ - \mathcal{P}\mathbf{q}_\Pi + \mathcal{P}\mathbf{q}_\Pi - \mathbf{q}_\Pi\|_{\mathbf{z}} \leq \|\mathbf{q}_+ - \mathcal{P}\mathbf{q}_\Pi\|_{\mathbf{z}} + \|\mathcal{P}\mathbf{q}_\Pi - \mathbf{q}_\Pi\|_{\mathbf{z}}$. Next notice that $\|\mathbf{q}_+ - \mathcal{P}\mathbf{q}_\Pi\|_{\mathbf{z}} = \|\mathcal{P}\mathcal{O}\mathbf{q}_+ - \mathcal{P}\mathbf{q}_\Pi\|_{\mathbf{z}} \leq \|\mathcal{O}\mathbf{q}_+ - \mathbf{q}_\Pi\|_{\mathbf{z}} = \|\mathcal{O}\mathbf{q}_+ - \mathcal{O}\mathbf{q}_\Pi\|_{\mathbf{z}} \leq \gamma\|\mathbf{q}_+ - \mathbf{q}_\Pi\|_{\mathbf{z}}$ by Lemma 2. Thus $(1-\gamma)\|\mathbf{q}_+ - \mathbf{q}_\Pi\|_{\mathbf{z}} \leq \|\mathcal{P}\mathbf{q}_\Pi - \mathbf{q}_\Pi\|_{\mathbf{z}}$. ∎

Linear function approximation in the dual case is a bit more complicated because here we are representing matrices, 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.

To begin, we assume the dual matrix H can be represented as a linear combination of basis matrices Υ and Γ, such that $H = \Upsilon W \Gamma$, where Υ is a fixed $|S||A| \times k$ matrix of row normalized basis distributions, Γ is a fixed $k \times |S||A|$ matrix of row normalized basis distributions, and W is a $k \times k$, row normalized matrix of adjustable weights. It is easy to verify that the constraints $\Upsilon, W, \Gamma \geq 0$ and $\Upsilon\mathbf{1} = \mathbf{1}$, $W\mathbf{1} = \mathbf{1}$, $\Gamma\mathbf{1} = \mathbf{1}$ implies $H \geq 0$ and $H\mathbf{1} = \mathbf{1}$. Thus the space of representable matrices is a k^2 dimensional simplex spanned by the two sets of basis distributions Υ and Γ.

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, we need to construct a projection operator 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 however, to ensure convergence. Here the pseudo-norm $\|\cdot\|_{\mathbf{z},\mathbf{r}}$ defined in Section 5.1 suits this purpose. Define the weighted projection operator $\mathcal{P}$ over matrices

$$\mathcal{P}H \quad = \quad \operatorname*{argmin}_{H' \in \text{simplex}(\Upsilon,\Gamma)} \|H - H'\|_{\mathbf{z},\mathbf{r}}$$

Owing to the constraints, this projection has to be computed by a quadratic program, rather than just simply solving a linear system

$$\hat{W} \quad = \quad \operatorname{argmin} W \geq 0, W\mathbf{1} = \mathbf{1} \|(H - \Upsilon W \Gamma)\mathbf{r}\|_{\mathbf{z}}{}^2$$

and then forming the projected matrix $\hat{H} = \Upsilon \hat{W} \Gamma$. A key result is that this projection operator is a non-expansion with respect to the pseudo-norm $\|\cdot\|_{\mathbf{z},\mathbf{r}}$.

Theorem 1 $\|\mathcal{P}H\|_{\mathbf{z},\mathbf{r}} \leq \|H\|_{\mathbf{z},\mathbf{r}}$

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(Υ, Γ), then onto the subspace of row normalized matrices span$(\Upsilon, \Gamma) \cap \{H : H\mathbf{1} = \mathbf{1}\}$, and finally onto the space of nonnegative matrices span$(\Upsilon, \Gamma) \cap \{H : H\mathbf{1} = \mathbf{1}\} \cap \{H : H \geq 0\}$. 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 $\mathcal{P}_1$.

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

where the third equality follows from Lemma 7. Since the overall projection is just a composition of non-expansions, it too must be a non-expansion. ∎

As in the primal, we can implement approximate dynamic programming 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, $\mathcal{P}\mathcal{O}$ 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 $\mathcal{P}\mathcal{O}$ by some representative $H_+ = \mathcal{P}\mathcal{O}H_+$.

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 2 $\|H_+ - H_\Pi\|_{\mathbf{z},\mathbf{r}} \leq \frac{1}{1-\gamma}\|H_+ - \mathcal{P}H_\Pi\|_{\mathbf{z},\mathbf{r}}$

Proof: In fact, the proof follows identical steps to the proof of Lemma 8, using the pertinent pseudo-norm and projection operators defined for the dual. ∎

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 $(1-\gamma)(\mathcal{P}\mathcal{O}\mathbf{q}) = (\mathcal{P}\mathcal{O}H)\mathbf{r}$. 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, then, the dual updates cannot diverge, even using compositions with the max operator $\mathcal{P}\mathcal{M}$; yet this update potentially diverges in the primal. For convergent compositions, like $\mathcal{P}\mathcal{O}$, even though the bound is not stronger, below we tend to see smaller approximation errors in the dual.

7. Gradient operators

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. Moreover, full least squares projections are usually not practical to compute either. The main intermediate step toward practical algorithms is to formulate gradient step operators that only approximate complete projections. Conveniently, gradient update and projection operators are independent of the on-policy and max operators and can be applied in either case. However, as we will see below, the gradient update operator causes significant instability in the max-policy update, to the degree that divergence is a common phenomenon (much more so than with full projections). Composing approximation with max operators 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 max updates is one of the most common algorithms attempted in reinforcement learning (Q-learning with function approximation), while also being the most unstable.

Gradient step updates are easily derived from a given projection operator. In this case, one always works directly with weight vectors $\mathbf{w}$ and weight matrices W, rather than complete $\mathbf{q}$ vectors or H matrices. For the primal case, the projection operator is equivalent to solving for a vector $\mathbf{w}$ of basis combination weights that minimizes the least square objective $\frac{1}{2}\|\Phi\mathbf{w} - \mathbf{q}_{targ}\|_{\mathbf{z}}^2$ The gradient of the objective is

$$\nabla\mathbf{w} \;=\; \Phi^\top(\Phi\mathbf{w} - \mathbf{q}_{targ}) \;=\; \Phi^\top(\mathbf{q} - \mathbf{q}_{targ})$$

The gradient operator can be defined with respect to a fixed step size α by

$$\mathcal{G}\mathbf{q} \;=\; \Phi(\mathbf{w} - \alpha\nabla\mathbf{w}) \;=\; \mathbf{q} - \alpha\Phi\Phi^\top(\mathbf{q} - \mathbf{q}_{targ})$$

The target vector $\mathbf{q}_{targ}$ is determined by the underlying dynamic programming update, so for example

$$\begin{aligned}\mathcal{G}\mathcal{O}\mathbf{q} &= \mathbf{q} - \alpha\Phi\Phi^\top(\mathbf{q} - \mathcal{O}\mathbf{q})\\ \mathcal{G}\mathcal{M}\mathbf{q} &= \mathbf{q} - \alpha\Phi\Phi^\top(\mathbf{q} - \mathcal{M}\mathbf{q})\end{aligned}$$

In our experiments below, the former always converges, whereas the latter diverges in at least half of our experiments.

In a real reinforcement learning scenario, these gradient update operators are applied pointwise with a single sampled transition, in place of a full expectation with respect to $P\Pi$, yielding the more familiar looking update

$$\begin{aligned}\nabla\mathbf{w} &= \Phi_{(sa,:)}^\top(\mathbf{q}_{(sa)} - \mathbf{q}_{targ\,(sa)})\\ \mathcal{G}\mathbf{q}_{(sa)} &= \Phi_{(sa,:)}(\mathbf{w} - \alpha\nabla\mathbf{w})\end{aligned}$$

In the dual representation, one can derive a gradient update operator in a similar way, except that it is important to maintain the constraints on the weight parameters W. As in the primal case, we start by considering the projection objective

$$\frac{1}{2}\|\Upsilon W\Gamma - H_{targ}\|_{\mathbf{z},\mathbf{r}}^2 \quad \text{s.t.} \quad W \geq 0, W\mathbf{1} = \mathbf{1}$$

The unconstrained gradient can be derived to be

$$\begin{aligned}\nabla W &= \Upsilon^\top Z\Upsilon W\Gamma\mathbf{r}\mathbf{r}^\top\Gamma^\top - \Upsilon^\top Z H_{targ}\mathbf{r}\mathbf{r}^\top\Gamma^\top\\ &= \Upsilon^\top Z(H - H_{targ})\mathbf{r}\mathbf{r}^\top\Gamma^\top\end{aligned}$$

However, this gradient step cannot be followed directly because we need to maintain the constraints. The conttraint $W\mathbf{1} = \mathbf{1}$ can be maintained by first projecting the gradient onto it, obtaining

$$\Delta W \;=\; \nabla W - \tfrac{1}{k}\nabla W\mathbf{1}\mathbf{1}^\top$$

which satisfies $\Delta W\mathbf{1} = 0$. The gradient operator can then be defined by

$$\mathcal{G}H \;=\; \Upsilon(W - \alpha^*\Delta W)\Gamma$$

where the step size α^* is possibly reduced to $\alpha^* \leq \alpha$ to maintain $W - \alpha^*\Delta W \geq 0$. The pointwise sample version of this update, suitable for RL problems, can be derived from the gradient update, but we do not have space to explore the resulting algorithm.

8. Experimental results

To investigate the effectiveness of the dual representations, we conducted experiments on randomly synthesized MDPs, on the *star problem*, and on the *mountain car* problem. The star problem is perhaps the most-cited example of a problem where Q-learning with function approximation (Baird, 1995) diverges, and the mountain car domain has been prone to divergence with some primal representations as well (Boyan & Moore, 1995). We too have observed the divergence of state-action value functions when using gradient-based updates with the max operator.

For the synthesized MDPs, we generated the dynamics and reward function of the MDPs randomly. We also choose random basis functions and basis distributions for projection since our goal is to investigate the convergence of the algorithms without carefully crafting features. We observed consistent convergence using dual representations with various random problems with different states, actions and bases. Here we only reported the plots of random MDPs with 100 states, 5 actions, and 10 bases, averaging over 500 repeats.

The star problem has 7 states and 2 actions. Baird stated that Q-learning with linear function approximation can diverge on this problem even when training on a fixed stochastic policy. We observed divergence in the primal case when using gradient updates, but all dual representation methods converged. The mountain car domain has continuous state and action spaces, which we discretize with a simple grid, resulting in an MDP with 222 states. In the primal representations with function approximation, we randomly generated basis functions. In the dual representations, we randomly picked the basis distributions.

For each problem domain, we set a finite horizon T. For on-policy algorithms, we measure the difference

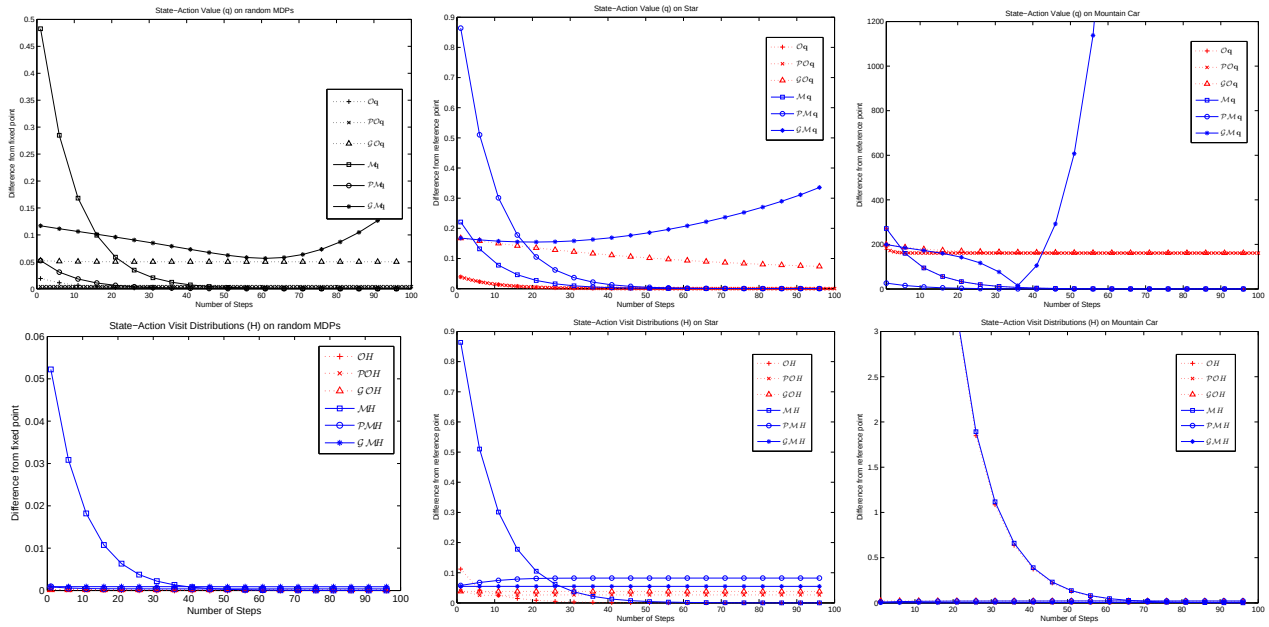

Figure 1. Behavior of the various update operators on different problems. For the $\mathcal{O}$ operator, plots show the distance from $\mathbf{q}$ or H to the fixed point determined by the policy. For the $\mathcal{M}$ operator, plots show the distance from the current state-action value $\mathbf{q}$ (either explicitly represented or implied by H) to the optimal function $\mathbf{q}^*$.

between the values generated by the algorithms and those generated by the fixed-point distribution. For max-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 gradient updates was 0.1 for primal representations and 10000 for dual representations. The initial values of state-action value functions (q) and state-action visit distributions (H) are chosen randomly. The discount factor was set to $\gamma = 0.9$. In all cases, algorithms using dual representations converged.

9. 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 demonstrated how this approach can be scaled up via 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, and the contribution of a novel perspective that yields new forms of prior knowledge that can be exploited in large domains.

References

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. *Proc. ICML*.

Bertsekas, D. (1995). *Dynamic programming and optimal control*, vol. 2. Athena Scientific.

Bertsekas, D., & Tsitsiklis, J. (1996). *Neuro-dynamic programming*. Athena Scientific.

Boyan, J. A., & Moore, A. W. (1995). Generalization in reinforcement learning: Safely approximating the value function. *NIPS 7* (pp. 369–376).

de Farias, D., & Van Roy, B. (2000). On the existence of fixed points for approximate value iteration and temporal-difference learning. *J. Optimization Theory and Applic.*, *105*, 589–608.

Puterman, M. (1994). *Markov decision processes: Discrete dynamic programming*. Wiley.

Ross, S. (1997). *Introduction to probability models*. Academic Press. 6th edition.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. MIT Press.

Tsitsiklis, J., & Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Trans. Automat. Control*, *42*, 674–690.

Wang, T., Bowling, M., & Schuurmans, D. (2007). Dual representations for dynamic programming and reinforcement learning. *Proceedings IEEE ADPRL*.