The Role of Lookahead and Approximate Policy Evaluation in Policy Iteration with Linear Value Function Approximation

Anna Winnicki, Joseph Lubars, Michael Livesay, R. Srikant

1 ECE and CSL, University of Illinois at Urbana-Champaign
2 Sandia National Laboratories

annaw5@illinois.edu, lubars2@illinois.edu, mlivesa@sandia.gov, rsrikant@illinois.edu

Abstract

When the sizes of the state and action spaces are large, solving MDPs can be computationally prohibitive even if the probability transition matrix is known. So in practice, a number of techniques are used to approximately solve the dynamic programming problem, including lookahead, approximate policy evaluation using an $m$-step return, and function approximation. In a recent paper, (Efroni et al. 2019) studied the impact of lookahead on the convergence rate of approximate dynamic programming. In this paper, we show that these convergence results change dramatically when function approximation is used in conjunction with lookahead and approximate policy evaluation using an $m$-step return. Specifically, we show that when linear function approximation is used to represent the value function, a certain minimum amount of lookahead and multi-step return is needed for the algorithm to even converge. And when this condition is met, we characterize the finite-time performance of policies obtained using such approximate policy iteration. Our results are presented for two different procedures to compute the function approximation: linear least-squares regression and gradient descent.

1 Introduction

Policy iteration and variants of policy iteration (Bertsekas 2019, 2011, Bertsekas and Tsitsiklis 1996) that solve dynamic programming problems rely on computations that are infeasible due to the sizes of the state and action spaces in modern reinforcement learning problems. As a remedy to this “curse of dimensionality,” several state-of-the-art algorithms (Silver et al. 2017a,b, Mnih et al. 2016) employ return for policy evaluation, lookahead using tree search, function approximation, and gradient descent to compute the function approximation, see (Section 2 and Bertsekas 2019) for a definition of these terms.

The recent work in (Efroni et al. 2019) considers a variant of policy iteration that utilizes lookahead and approximate policy evaluation using an $m$-step return (see Section 2 for definitions of these terms). As stated in the motivation in (Efroni et al. 2019), we note that lookahead can be approximated well in practice using Monte Carlo Tree Search (MCTS) (Kocsis and Szepesvári 2006, Browne et al. 2012) even though in theory, it has exponential complexity (Shah, Xie, and Xu 2020). Motivated by policy iteration, the algorithm in (Efroni et al. 2019) estimates the value function associated with a policy and aims to improve the policy at each step. Policy improvement is achieved by obtaining the “greedy” policy in the case of policy iteration or a lookahead policy in the work of (Efroni et al. 2019), which involves applying the Bellman operator several times to the current iterate before obtaining the greedy policy. The idea is that the application of the Bellman operator several times gives a more accurate estimate of the optimal value function.

Then, similar to policy iteration, the algorithm in (Efroni et al. 2019) aims to evaluate the new policy. The algorithm in (Efroni et al. 2019) uses an $m$-step return to compute the value function associated with a policy, i.e., it applies the Bellman operator associated with the policy $m$ times.

The work of (Efroni et al. 2019) establishes that a lookahead can significantly improve the rate of convergence if one uses the value function computed using lookahead in the approximate policy evaluation step. Our main interest is understanding how these convergence results change when the state-space is very large and one has to resort to function approximation of the value function. Our contributions are as follows:

1. We examine the impact of lookahead and $m$-step return on approximate policy iteration with linear function approximation. We do not need to evaluate an approximate value function for all the states at each iteration but can simply evaluate it at some states. Since we use function approximation, we need different proof techniques than in (Efroni et al. 2019) and one consequence of this is that the finite-time performance bounds that we obtain for the algorithm (see Appendix A) require that the sum of lookahead and the number of steps in the $m$-step return is sufficiently large. We demonstrate through an extension of a counter example in (Tsitsiklis and van Roy 1994) that such a condition is necessary for convergence with function approximation unlike the tabular setting in (Efroni et al. 2019).

2. Our first results mentioned above assume that one solves a least-squares problem at each iteration to obtain the weights associated with the feature vectors in the function approximation of the value function. We then obtain finite-time performance bounds for a more practical and widely-used scheme where one step of gradient descent is used to update the weights of the value function approximation at each iteration. The finite-time performance bounds for the gradient descent algorithm can be
found in Section [4]

3. The sufficient condition on the minimum amount of lookahead and return for convergence does not depend on the size of the state space but depends on the feature vectors used for function approximation. Additionally, our results indicate that the approximation error of our algorithm can be improved if the feature vectors can represent the value function well. While neural networks are not linear function approximators, recent results motivated by the NTK (neural tangent kernel) analysis of neural networks suggest that they can be approximated as linear combinations of basis functions [Jacot, Gabriel, and Hongler 2018; Cao and Gu 2019]. Thus, our results can potentially shed light on why the combination of the representation capability of neural networks and tree-search methods work well in practice, although further work is necessary to make this connection precise.

4. We extend our results to the case where the m-step return is replaced by an approximation of TD-learning, as in the work of (Efroni et al. 2019). These results can be found in Appendix [F] and Appendix [G].

5. We complement our theoretical results with experiments on the same grid world problem as in (Efroni et al. 2019). These experiments are presented in a later section and in Appendix [H].

In addition to the work of (Efroni et al. 2019), there is a long history of other work on approximate policy iteration. However, these other works typically do not model the tree search processes (lookahead and m-step return) in conjunction with function approximation and gradient descent as we have done in this paper. We will compare our results to some of these prior works in a later section.

2 Preliminaries

We consider a Markov Decision Process (MDP), which is defined to be a 5-tuple \((S, A, P, R, \alpha)\). The finite set of states of the MDP is \(S\). There exists a finite set of actions associated with the MDP \(A\). Let \(P_{ij}(a)\) be the probability of transitioning from state \(i\) to state \(j\) when taking action \(a\) \(\in A\). We denote by \(s_k\) the state of the MDP and by \(a_k\) the corresponding action at time \(k\). We associate with state \(s_k\) and action \(a_k\) a non-deterministic reward \(r(s_k, a_k) \in [0, 1]\forall s_k \in S, a_k \in A\). We assume that the rewards are uniformly bounded. Our objective is to maximize the cumulative discounted reward with discount factor \(\alpha \in (0, 1)\), \(\sum_{i=0}^{\infty} \alpha^i r(s_k, a_k)\).

Towards this end, we associate with each state \(s \in S\) a deterministic policy \(\mu(s) \in A\) which prescribes an action to take. For every policy \(\mu\) and every state \(s \in S\) we define \(J^\mu(s)\) as follows:

\[J^\mu(s) := E\left[\sum_{i=0}^{\infty} \alpha^i r(s_k, \mu(s_k))\right].\]

We define the optimal reward-to-go \(J^*\) as \(J^*(s) := \min_{\mu} J^\mu(s)\). The objective is to find a policy \(\mu\) that maximizes \(J^\mu(s)\) for all \(s \in S\). Towards the objective, we associate with each policy \(\mu\) a function \(T^\mu : \mathbb{R}^n \rightarrow \mathbb{R}^n\) where for \(J \in \mathbb{R}^n\), the \(s\)th component of \(T^\mu J\) is

\[(T^\mu J)(s) = r(s, \mu(s)) + \alpha \sum_{j=1}^{\vert S \vert} p_{kj}(\mu(s))J(j),\]

for all \(s \in S\). If function \(T^\mu\) is applied \(m\) times to vector \(J \in \mathbb{R}^{\vert S \vert}\), we call the result \(T^mJ\). We say that \(T^*\) is the \(m\)-return corresponding to \(J\) or the “return,” when \(J\) and \(m\) are understood.

Similarly, we define the Bellman operator \(T : \mathbb{R}^{\vert S \vert} \rightarrow \mathbb{R}^{\vert S \vert}\) with the \(s\)th component of \(TJ\) being

\[(TJ)(s) = \max_{a} \left\{r(s, a) + \alpha \sum_{j=1}^{\vert S \vert} p_{sj}(a)J(j)\right\}. \tag{1}\]

The policy corresponding to the \(T\) operator is defined as the greedy policy. If function \(T\) is applied \(H\) times to vector \(J \in \mathbb{R}^{\vert S \vert}\), we call the result \(T^HJ\) - the \(H\)-step “lookahead” corresponding to \(J\). The greedy policy corresponding to \(T^H\) is called the \(H\)-step lookahead policy, or the lookahead policy, when \(H\) is understood. More precisely, given an estimate \(J\) of the value function, the lookahead policy is the policy \(\mu\) such that \(T^\mu(T^{H-1}J) = T^HJ\).

It is well known that each time the Bellman operator is applied to a vector \(J\) to obtain \(TJ\), the following holds:

\[\|TJ - J^*\|_\infty \leq \alpha \|J - J^*\|_\infty.\]

Thus, applying \(T\) to obtain \(TJ\) gives a better estimate of the value function than \(J\).

The Bellman equations state that the vector \(J^\mu\) is the unique solution to the linear equation

\[J^\mu = T^\mu J^\mu. \tag{2}\]

Additionally, we have that \(J^*\) is a solution to

\[J^* = T^*J^*.\]

Note that every greedy policy w.r.t. \(J^*\) is optimal and vice versa [Bertsekas and Tsitsiklis 1996].

We will now state several useful properties of the operators \(T\) and \(T^\mu\). Consider the vector \(e \in \mathbb{R}^{\vert S \vert}\) where \(e(i) = 1\forall i \in 1, 2, \ldots, \vert S \vert\). We have:

\[T(J + ce) = TJ + cJe, \quad T^\mu(J + ce) = T^\mu J + cJe. \tag{3}\]

Operators \(T\) and \(T^\mu\) are also monotone:

\[J \leq J' \implies TJ \leq TJ', \quad T^\mu J \leq T^\mu J'. \tag{4}\]

3 Least Squares Function Approximation Algorithm

Our algorithm is described in Algorithm [1]. The algorithm is an approximation to policy iteration. At each iteration index, say, \(k\), we have an estimate of the value function, which we denote by \(J_k\). To obtain \(J_{k+1}\), we perform a lookahead to improve the value function estimate at a certain number of states (denoted by \(D_k\)) which can vary with each iteration. For example, \(D_k\) could be chosen as the states visited
Algorithm 1: Least Squares Function Approximation Algorithm

**Input:** $J_0$, $m$, $H$, feature vectors $\{\phi(i)\}_{i \in S}, \phi(i) \in \mathbb{R}^d$ and subsets $D_k \subseteq S$, $k = 0, 1, \ldots$. Here $D_k$ is the set of states at which we evaluate the current policy at iteration $k$.

1. Let $k = 0$.
2. Let $\mu_{k+1}$ be such that $\|T^H J_k - T_{\mu_{k+1}} T^{H-1} J_k\|_\infty \leq \varepsilon$.
3. Compute $\hat{J}_{\mu_{k+1}}(i) = T^m_{\mu_{k+1}} T^{H-1} (J_k)(i) + w_{k+1}(i)$ for $i \in D_k$.
4. Choose $\theta_{k+1}$ to solve
   \[
   \min_\theta \sum_{i \in D_k} (\Phi \theta(i) - \hat{J}_{\mu_{k+1}}(i))^2, \]
   where $\Phi$ is a matrix whose rows are the feature vectors.
5. $J_{k+1} = \Phi \theta_{k+1}$.
6. Set $k \leftarrow k + 1$. Go to 2.

when performing a tree search to approximate the lookahead process. During the lookahead process, we note that we will also obtain a $H$-step lookahead policy, which we denote by $\mu_{k+1}$. As noted in the Introduction, the computation of $T^{H-1}(J_k)(i)$ for $i \in D_k$ in Step 3 of Algorithm 1 may be computationally infeasible; however, as noted in [Efroni et al., 2019], techniques such as Monte Carlo tree search (MCTS) are employed in practice to approximately estimate $T^{H-1}(J_k)(i)$.

We obtain estimates of $J_{\mu_{k+1}}(i)$ for $i \in D_k$, which we call $\hat{J}_{\mu_{k+1}}(i)$. To obtain the estimate of the $J_{\mu_{k+1}}(i)$, we perform a $m$-step return with policy $\mu_{k+1}$, and obtain the estimate of $T^m_{\mu_{k+1}} T^{H-1} J_k(i)$ for $i \in D_k$. We model the approximation errors in lookahead and return by adding noise to the output of these steps. In order to estimate the value function for states not in $D_k$, we associate with each state $i \in S$ a feature vector $\phi(i) \in \mathbb{R}^d$ where typically $d << |S|$. The matrix comprised of the feature vectors as rows is denoted by $\Phi$. We obtain estimates of $T^m_{\mu_{k+1}} T^{H-1} J_k$ for states in $D_k$ and use those estimates to find the best fitting $\theta \in \mathbb{R}^d$, i.e.,

\[
\min_\theta \sum_{i \in D_k} (\Phi \theta(i) - \hat{J}_{\mu_{k+1}}(i))^2, \]

Our algorithm then computes $\theta_{k+1}$. It uses the $\theta_{k+1}$ to obtain $J_{k+1} = \Phi \theta_{k+1}$. The process then repeats. Note that to estimate the value function associated with policy $\mu_{k+1}$, we compute $T^m_{\mu_{k+1}} T^{H-1} J_k(i)$ for $i \in D_k$. Another alternative is to instead compute $T^m_{\mu_{k+1}} J_k(i)$. It was shown in [Efroni et al., 2019] that the former option is preferable because it has a certain contraction property. Thus, we have chosen to use this computation in our algorithm as well. However, we show in Appendix B that the algorithm also converges with the second option if $m$ is chosen to be sufficiently large.

To analyze the above algorithm, we make the following assumption which states that we explore sufficient number of states during the policy evaluation phase at each iteration.

**Assumption 1.** For each $k \geq 0$, rank $\{\phi(i)\}_{i \in D_k} = d$.

We assume that the noise, $w_{k+1}$, is bounded.

**Assumption 2.** For some $\varepsilon' > 0$, $\|w_k\|_\infty \leq \varepsilon'$.

We also assume that the rewards are bounded.

**Assumption 3.** $r(i, u) \in [0, 1]$ for all $i, u$.

Using Assumption 1, $J_{k+1}$ can be written as

\[
J_{k+1} = \Phi \theta_{k+1} = \Phi (\Phi_D^\top \Phi_D)^{-1} \Phi_D^\top \mathcal{P}_k \hat{J}_{\mu_{k+1}}, \]

where $\Phi_D$ is a matrix whose rows are the feature vectors of the states in $D_k$ and $\mathcal{P}_k$ is a matrix of zeros and ones such that $\mathcal{P}_k J_{\mu_{k+1}}$ is a vector whose elements are a subset of the elements of $J_{\mu_{k+1}}$ corresponding to $D_k$.

Written concisely, our algorithm is as follows:

\[
J_{k+1} = M_{k+1}(T^m_{\mu_{k+1}} T^{H-1} J_k + w_k), \]

where $M_{k+1}$ is defined in step 2 of the algorithm. Note that $w_k(i)$ for $i \notin D_k$ does not affect the algorithm, so for convenience we can define $w_k(i) = 0$ for $i \notin D_k$.

Now we will state our theorem which characterizes the role of lookahead ($H$) and return ($m$) on the convergence of approximate policy iteration with function approximation.

**Theorem 1.** Suppose that $m$ and $H$ satisfy $m + H - 1 > \log(\delta_1)/\log(1/\alpha)$, where

\[
\delta_1 := \sup_k \|M_k\|_\infty = \sup_k \|\Phi (\Phi_D^\top \Phi_D)^{-1} \Phi_D^\top \mathcal{P}_k\|_\infty, \]

then

\[
\limsup_{k \to \infty} \|J^m_{\mu_k} - J^*\|_\infty \leq \frac{c_{m,H}}{(1 - \alpha)(1 - \alpha^{H-1})}, \]

where

\[
c_{m,H} := 2\alpha^H \left( \frac{\alpha^m - \alpha^{m+H-1}}{1 - \alpha} \delta_2 + \delta_1 + \delta_1 + \frac{1}{1 - \alpha} \right) + \|J_0\|_\infty + \varepsilon, \]

and $\delta_2 := \sup_{k, \mu_k} \|M_k J^m_{\mu_k} - J^m_{\mu_k}\|_\infty$.

The proof of Theorem 1 is subsumed by the proof of Theorem 2 (presented in the next section) with minor modifications. Therefore, we will outline the modifications after the proof of Theorem 1 and in Appendix A. It is important to note that while the bounds given in Theorem 1 and Theorem 2 are asymptotic, finite-time bounds are a byproduct of our proofs and can be found in Appendix A and Section 4 respectively.

We additionally characterize the approximation error of our iterates, $J_k$, by computing bounds on the asymptotic error $\limsup_{k \to \infty} \|J_k - J^*\|_\infty$. The bounds along with their derivations can be found in Appendix C. The corresponding finite-time bounds can be easily obtained from the proof of Proposition 1 in Appendix C.
It is important to note that the upper bounds on $||J_{\mu} - J^*||_{\infty}$ and $||J_k - J^*||_{\infty}$ illustrate that the convergence rate of $||J_{\mu} - J^*||_{\infty}$ is much faster than the convergence rate of $||J_k - J^*||_{\infty}$. Thus, algorithms need not wait for the value function estimates to converge before the corresponding greedy policy reaches near optimality.

In (Bertsekas 2021), it is noted that, in reinforcement learning to play computer games or board games, it is not uncommon during training to get a relatively crude estimate of the value function, which is improved by lookahead and $m$-step return during actual game play. Our analysis would also apply to this situation—we have not explicitly differentiated between training and game play in our analysis.

Theorem 1 can be used to make the following observation: how close $J_{\mu}$ is to $J^*$ depends on four factors—the representation power of the feature vectors and the feature vectors themselves ($\delta_2, \delta_1$), the amount of lookahead ($H$), the extent of the return ($m$) and the approximation in the policy determination and policy evaluation steps ($\varepsilon$ and $\varepsilon'$). Further, by examining $c_m,H$ one can see that lookahead and return help mitigate the effect of feature vectors and their ability to represent the value functions. We note that (Efroni et al. 2019) also consider a version of the algorithm where in Step 3 $T_{\mu}^m T^H J_k$ is replaced by $T_{\mu}^m J_k$. We obtain a performance bound for this algorithm with function approximation in Appendix B under the assumption that $m$ is large.

### 3.1 Counter-Example

Even though, in practice, $J_{\mu}$ is what we are interested in, the values $J_k$ computed as part of our algorithm should not go to $\infty$ since the algorithm would be numerically unstable otherwise. In Appendix C we provide a bound on $||J_k - J^*||_{\infty}$ when $m + H - 1$ is sufficiently large as in Theorem 1. In this subsection, we show that, when this condition is not satisfied, $J_k$ can become unbounded.

The example we use is depicted in Figure 1. There are two policies, $\mu^a$ and $\mu^b$ and the transitions are deterministic under the two policies. The rewards are deterministic and only depend on the states. The rewards associated with states are denoted by $r(x_1)$ and $r(x_2)$, with $r(x_1) > r(x_2)$. Thus, the optimal policy is $\mu^a$. We assume scalar features $\phi(x_1) = 1$ and $\phi(x_2) = 2$.

We fix $H = 1$. The MDP follows policy $\mu^a$ when:

\[ J_k(x_1) > J_k(x_2) \implies \theta_k > 2 \theta_k. \]

Thus, as long as $\theta_k > 0$, the lookahead policy will be $\mu^b$.

We will now show that $\theta_k$ increases at each iteration when $\delta_1 a_m^{m+H-1} > 1$. We assume that $\theta_0 > 0$ and $D_k = \{1, 2\}$ for all $k$. A straightforward computation shows that $\delta_1 = \frac{a_1}{a_2}$.

At iteration $k+1$, suppose $\mu_{k+1} = \mu^b$, our $J_{\mu_{k+1}}(i)$ for $i = 1, 2$ are as follows:

\[ J_{\mu_{k+1}}(1) = r(x_1) + \sum_{i=1}^{m-1} r(x_1) x^i + 2 a_m \theta_k \]

\[ J_{\mu_{k+1}}(2) = r(x_2) + \sum_{i=1}^{m-1} r(x_2) x^i + 2 a_m \theta_k. \]

Thus, from Step 4 of Algorithm 1:

\[ \theta_{k+1} = \arg \min_{\theta} \sum_{i=1}^{2} \left( (\Phi \theta)(i) - J_{\mu_{k+1}}(i) \right)^2 \]

\[ = \sum_{i=1}^{2} \alpha^i r(x_1) + 2 \sum_{i=1}^{m-1} \alpha^i r(x_2) \]

\[ + \frac{6 a_m \theta_k}{5} \]

\[ = \theta_{k+1} > \frac{6}{5} a_m \theta_k. \]

Thus, since $\theta_0 > 0$ and $H = 1$, when $\frac{6}{5} a_{m+H-1} \theta_k = \delta_1 a_m^{m+H-1} > 1$, $\theta_k$ goes to $\infty$.

### 3.2 Numerical Results

In this section, we study the convergence behavior of our algorithm, including the impact of the choice of feature vectors. We present some of our experimental results; for our complete set of results, refer to Appendix C.

As in (Efroni et al. 2018a) and (Efroni et al. 2019), we assume a deterministic grid world problem played on an $N \times N$ grid. The states are the squares of the grid and the actions are {‘up’, ‘down’, ‘right’, ‘left’, and ‘stay’}, which move the agent in the prescribed direction, if possible. In each experiment, a goal state is chosen uniformly at random to have a reward of 1, while every other state has a fixed reward drawn uniformly from $[-0.1, 0.1]$. For our experiments, $N = 25$ and $\alpha = 0.9$. We implemented Algorithm 1 for two particular choices of feature vector:

1. Random feature vectors (dimension 20): each entry of the matrix $\Phi$ is an independent $N(0, 1)$ random variable.

2. Designed feature vectors: the feature vector for a state whose coordinates are $(x, y)$ is $(x, y, d)^T$, where $d$ is the number of steps required to reach the goal from state $(x, y)$.

For each choice of feature vectors, we varied $H$ and $m$ while keeping the other parameter constant at $H = 3$ or $m = 3$, and plotted the error $||J_k - J^*||_{\infty}$ for each iteration $k$, averaged over ten trials for each value of $H$ and $m$. As seen in Figure 2, although $J_k$ diverges for small values of $H$ and $m$, for sufficiently large values of $m + H$, the value function converges to a region around the optimal solution. Recall that our theorems suggest that the amount of lookahead and return depends on the choice of the feature vectors. Our experiments support this observation as well. The amount of lookahead and $m$-step return required is high (often over 30) for random feature vectors, but we are able to significantly reduce the amount required by using the designed feature vectors which better represent the states.

Figure 1: An example illustrating the necessity of the condition in Theorem 1.
γ, m, and J of \( \tilde{\theta} \). Thus, the value of \( J \) eventually stops diverging. (Bottom) For designed feature vectors, a smaller amount of lookahead and m-step return are needed to prevent \( J \) from diverging.

### 4 Gradient Descent Algorithm

Solving the least-squares problem in Algorithm [1] involves a matrix inversion, which can be computationally difficult if the dimension of the feature vectors is large. So we propose an alternative algorithm which performs one step of gradient descent at each iteration, where the gradient refers to the gradient of the least-squares objective in (4). The gradient descent-based algorithm is presented in Algorithm 2.

In order to present our main result for the gradient descent version of our algorithm, we define \( \theta^\mu \) for any policy \( \mu \) as follows:

\[
\theta^\mu := \arg\min_{\theta} \frac{1}{2} \| \Phi D_{\theta} \theta - \mathcal{P}_k J^\mu \|^2.
\]

In other words, \( \theta^\mu \) represents the function approximation of \( J^\mu \). We also define another quantity \( \hat{\theta}^\mu \) which will be used in the proof of the theorem:

\[
\hat{\theta}^\mu := \arg\min_{\theta} \frac{1}{2} \| \Phi D_{\theta} \theta - \mathcal{P}_k (T_{\mu k} T^{H-1} J_{k-1} + w_k) \|^2.
\]

Thus, \( \hat{\theta}^\mu \) represents the function approximation of the estimate of \( J^\mu \) obtained from the m-step rollout.

We now present our main result:

**Theorem 2.** Suppose that Assumptions 1, 2 hold and further, \( H, m, \) and \( J \) satisfy

\[
\alpha' + (1 + \alpha') \alpha^{m + H - 1} \delta_1' < 1,
\]

where

\[
\alpha' := \sup_{k} \| I - \gamma \Phi D_{\theta} \Phi D_{\theta} \|_2, \quad \text{and}
\]

\[
\delta_1' := \sup_{k} \| \Phi \|_{\infty} \left( \Phi \right)^{-1} \| \Phi \|_{\infty} \mathcal{P}_k \|_{\infty}.
\]

Then

\[
\lim_{k \to \infty} \sup_{k} \| J^\mu_k - J^* \|_{\infty} \leq \frac{c_{m,H,\gamma}}{(1 - \alpha)(1 - \alpha^{H-1})},
\]

where

\[
c_{m,H,\gamma} := 2 \alpha^H \| \Phi \|_{\infty} u_{m,H,\gamma} + \delta_2'
\]

and

\[
\delta_2' := \sup_{k,\mu_k} \| \Phi \theta^\mu_k - J^\mu_k \|_{\infty}, C := \frac{\sigma_{\min,\theta}}{1 - \alpha} + 2 \sigma_{\min,\theta} \delta_2'
\]

where

\[
\alpha' = (1 + \alpha') \left[ \alpha^m \frac{\delta_1'}{\| \Phi \|_{\infty}} + \alpha^{m - H + 1} \delta_1' + \varepsilon' \right] - \left( (1 + \alpha') \alpha^{m + H - 1} \delta_1' + \varepsilon \right) + \sup_{k,\mu_k} \| \theta^\mu_k \|_{\infty} + \| \theta_0 \|.
\]

Note that by choosing a step-size \( \gamma \) that is sufficiently small, one can ensure that the condition on \( \alpha' \) is satisfied.

**Proof.** The proof of Theorem 2 is somewhat involved, so we outline the steps of the proof before each step of the proof.

**Step 1**

1. \( k = 0, J_0 = \Phi \theta_0 \).
2. Let \( \mu_k + 1 \) be such that \( \| T^H J_k - T_{\mu_k + 1} T^H J_k \|_{\infty} \leq \varepsilon \).
3. Compute \( \hat{\theta}^\mu_k + 1 (i) = T_{\mu_k + 1} T^H J_k (i) + w_{k+1} (i) \) for \( i \in \mathcal{D}_k \).
4. \( \theta_{k+1} = \theta_k - \gamma \nabla_{\theta} c(\theta, \hat{\theta}^\mu_k + 1) \) (8)

where

\[
c(\theta, \hat{\theta}^\mu_k + 1) := \min_{\theta} \frac{1}{2} \sum_{i \in \mathcal{D}_k} \left( (\Phi \theta)(i) - \hat{\theta}^\mu_k + 1 (i) \right)^2.
\]

5. \( J_{k+1} = \Phi \theta_{k+1} \).
6. Set \( k \leftarrow k + 1 \). Go to 2.
In this step, since $\theta_{k+1}$ is obtained by taking a step of gradient descent towards $\tilde{\theta}^{\mu k+1}$ beginning from $\theta_k$, we will show the following property:

$$\left\| \theta_{k+1} - \tilde{\theta}^{\mu k+1} \right\|_\infty \leq \alpha' \left\| \theta_k - \tilde{\theta}^{\mu k+1} \right\|_\infty.$$

**Proof of Step 1:** Recall that the iterates in Equation (8) can be written as follows:

$$\theta_{k+1} = \theta_k - \nabla c(\theta; J^{\mu k+1})$$

$$= \theta_k - \nabla \left( \Phi_{Dk}^T \Phi_{Dk} \theta_k - \Phi_{Dk}^T \mathcal{P}_k(T_{\mu k+1}^m T^{H-1} J_k + w_k) \right).$$

Since

$$0 = \nabla c(\theta; J^{\mu k+1})|_{\theta_{k+1}}$$

$$= \Phi_{Dk}^T \Phi_{Dk} \tilde{\theta}^{\mu k+1} - \Phi_{Dk}^T \mathcal{P}_k(T_{\mu k+1}^m T^{H-1} J_k + w_k),$$

we have the following:

$$\theta_{k+1} = \theta_k - \nabla \left( \Phi_{Dk}^T \Phi_{Dk} \theta_k - \Phi_{Dk}^T \Phi_{Dk} \tilde{\theta}^{\mu k+1} 
- \Phi_{Dk}^T \mathcal{P}_k(T_{\mu k+1}^m T^{H-1} J_k + w_k) 
+ \Phi_{Dk}^T \mathcal{P}_k(T_{\mu k+1}^m T^{H-1} J_k + w_k) \right)$$

$$= \theta_k - \nabla \Phi_{Dk}^T \Phi_{Dk} (\theta_k - \tilde{\theta}^{\mu k+1}).$$

Subtracting $\tilde{\theta}^{\mu k+1}$ from both sides gives:

$$\theta_{k+1} - \tilde{\theta}^{\mu k+1} = \theta_k - \tilde{\theta}^{\mu k+1} - \nabla \Phi_{Dk}^T \Phi_{Dk} (\theta_k - \tilde{\theta}^{\mu k+1})$$

$$= (I - \nabla \Phi_{Dk}^T \Phi_{Dk}) (\theta_k - \tilde{\theta}^{\mu k+1}).$$

Thus,

$$\left\| \theta_{k+1} - \tilde{\theta}^{\mu k+1} \right\|_\infty$$

$$\leq \left\| (I - \nabla \Phi_{Dk}^T \Phi_{Dk}) (\theta_k - \tilde{\theta}^{\mu k+1}) \right\|_\infty$$

$$\leq \left\| (I - \nabla \Phi_{Dk}^T \Phi_{Dk}) \right\|_2 \left\| \theta_k - \tilde{\theta}^{\mu k+1} \right\|_\infty$$

$$\leq \sup_k \left\| I - \nabla \Phi_{Dk}^T \Phi_{Dk} \right\|_2 \left\| \theta_k - \tilde{\theta}^{\mu k+1} \right\|_\infty \leq \alpha' \left\| \theta_k - \tilde{\theta}^{\mu k+1} \right\|_\infty.$$

**Step 2** Using the previous step in conjunction with contraction properties of the Bellman operators, we will show the following:

$$\left\| \theta_k - \tilde{\theta}^{\mu k} \right\|_\infty$$

$$\leq \left( \alpha' + (1 + \alpha') \alpha^{m+H-1} \delta_1' \right) \left\| \theta_{k-1} - \tilde{\theta}^{\mu k-1} \right\|_\infty + \alpha'C$$

$$+ (1 + \alpha') \left[ \alpha^m \frac{\delta_1'}{|\Phi|_\infty} \left( \frac{1 + \alpha^{H-1}}{1 - \alpha} + \alpha^{H-1} \delta_2' \right) + \varepsilon' \right].$$

We will iterate over $k$ to get a bound on $\left\| \theta_k - \tilde{\theta}^{\mu k} \right\|_\infty$ that does not depend on $k$.

**Proof of Step 2**

We have the following:

$$\left\| \theta_k - \tilde{\theta}^{\mu k} \right\|_\infty \leq \left\| \theta_k - \tilde{\theta}^{\mu k} \right\|_\infty + \left\| \tilde{\theta}^{\mu k} - \theta^{\mu k} \right\|_\infty$$

$$\leq \alpha' \left\| \theta_{k-1} - \tilde{\theta}^{\mu k-1} \right\|_\infty + \alpha' \left\| \tilde{\theta}^{\mu k} - \theta^{\mu k} \right\|_\infty$$

$$+ \left\| \tilde{\theta}^{\mu k} - \theta^{\mu k} \right\|_\infty$$

$$\leq \alpha' \left\| \theta_{k-1} - \tilde{\theta}^{\mu k-1} \right\|_\infty + \alpha' \left\| \theta^{\mu k} - \tilde{\theta}^{\mu k-1} \right\|_\infty$$

$$+ (1 + \alpha') \left\| \tilde{\theta}^{\mu k} - \theta^{\mu k} \right\|_\infty \leq \left( \alpha' + (1 + \alpha') \alpha^{m+H-1} \delta_1' \right) \left\| \theta_{k-1} - \tilde{\theta}^{\mu k-1} \right\|_\infty + \alpha'C$$

$$+ (1 + \alpha') \left[ \alpha^m \frac{\delta_1'}{|\Phi|_\infty} \left( \frac{1 + \alpha^{H-1}}{1 - \alpha} + \alpha^{H-1} \delta_2' \right) + \varepsilon' \right].$$

Iterating over $k$, we get that for all $k$,

$$\left\| \theta_k - \tilde{\theta}^{\mu k} \right\|_\infty \leq \left( \alpha' + (1 + \alpha') \alpha^{m+H-1} \delta_1' \right) \left\| \theta_{k-1} - \tilde{\theta}^{\mu k-1} \right\|_\infty + \alpha'C$$

$$+ (1 + \alpha') \left[ \alpha^m \frac{\delta_1'}{|\Phi|_\infty} \left( \frac{1 + \alpha^{H-1}}{1 - \alpha} + \alpha^{H-1} \delta_2' \right) + \varepsilon' \right].$$

$$\leq \left( \alpha' + (1 + \alpha') \alpha^{m+H-1} \delta_1' \right) \left\| \theta_{k-1} - \tilde{\theta}^{\mu k-1} \right\|_\infty + \alpha'C$$

$$+ (1 + \alpha') \left[ \alpha^m \frac{\delta_1'}{|\Phi|_\infty} \left( \frac{1 + \alpha^{H-1}}{1 - \alpha} + \alpha^{H-1} \delta_2' \right) + \varepsilon' \right].$$

$$= \alpha' C + (1 + \alpha') \alpha^m \frac{\delta_1'}{|\Phi|_\infty} \left( \frac{1 + \alpha^{H-1}}{1 - \alpha} + \alpha^{H-1} \delta_2' \right) + \varepsilon'$$

$$\leq \left( 1 - (\alpha' + (1 + \alpha') \alpha^{m+H-1} \delta_1') \right) \alpha' C + (1 + \alpha') \alpha^m \frac{\delta_1'}{|\Phi|_\infty} \left( \frac{1 + \alpha^{H-1}}{1 - \alpha} + \alpha^{H-1} \delta_2' \right) + \varepsilon' .$$

$$= \alpha' C + (1 + \alpha') \alpha^m \frac{\delta_1'}{|\Phi|_\infty} \left( \frac{1 + \alpha^{H-1}}{1 - \alpha} + \alpha^{H-1} \delta_2' \right) + \varepsilon' .$$

$$+ \sup_{k, \mu k} \| \tilde{\theta}^{\mu k} \|_\infty + \| \theta_0 \|. \tag{11}$$
where we have used the fact that $\alpha^r + (1 + \alpha^r) \alpha^{m+H-1} \delta_1 < 1$ from the assumption in the statement of Theorem 2 and the fact that $\|J^0\|_\infty \leq 1/(1 - \alpha)$ due to Assumption 2.

**Step 3**

Since $J_k = \Phi \theta_k$ and $\Phi \theta^\mu_k$ is the best approximation in $\mathbb{R}^d$ of $J^\mu$, we will show the following bounds:

$$\|J^\mu_k - J_k\|_\infty \leq \|\Phi\|_\infty \|\theta_k - \theta^\mu_k\|_\infty + \delta_2'. $$

Using the previous step, we will get a bound on $\|J^\mu_k - J_k\|_\infty$ that does not depend on $k$.

**Proof of Step 3**

$$\|J^\mu_k - J_k\|_\infty = \|\Phi \theta_k - J^\mu_k\|_\infty \leq \|\Phi\|_\infty \|\theta_k - \theta^\mu_k\|_\infty + \delta_2'. $$

(12) Taking the limit as $j \to \infty$ on both sides, we have the following:

$$J^\mu_{k+1} - J_k \leq \frac{c_{m,H,\gamma} e}{1 - \alpha} $$

The rest of proof is straightforward. Subtract $J^\star$ from both sides of the previous inequality and using the contraction property of $T$, we get

$$J^\mu_{k+1} - J^\star \leq \frac{c_{m,H,\gamma} e}{1 - \alpha} $$

which is equivalent to

$$\|J^\mu_{k+1} - J^\star\|_\infty \leq \alpha^{H-1} \|J^\mu_k - J^\star\|_\infty e + \frac{c_{m,H,\gamma} e}{1 - \alpha},$$

(14) since $J^\star \leq J^\mu$ for all policies $\mu$. Iterating over $k$, we get Theorem 2.

**Step 4**

We will establish the following bound on $T^H_{\mu_{k+1}} J^\mu_k$ using the contraction property of Bellman operators and property in (3):

$$T^H_{\mu_{k+1}} J^\mu_k \leq \alpha^H \|J^\mu_k - J_k\|_\infty e + T^H_{\mu_{k+1}} J^\mu_k e e$$

where the last inequality above follows from the definition of $p_{m,H,\gamma}$ in the previous step. Using properties in (3) and monotonicity, we will repeatedly apply $T^H_{\mu_{k+1}}$ to both sides and take limits to obtain the following:

$$J^\mu_{k+1} - J^\star \leq \frac{c_{m,H,\gamma} e}{1 - \alpha} $$

The bound in equation (14) gives us the following:

$$\|J^\mu_k - J^\star\|_\infty \leq \alpha^{H-1} \|J^\mu_k - J^\star\|_\infty e + \frac{c_{m,H,\gamma} e}{1 - \alpha},$$

for $k > 0$.

The rest of the proof is similar to the arguments in (Bertsekas and Tsitsiklis [1996]). (Bertsekas [2019]), however, we have to explicitly incorporate the role of lookahead ($H$) in the remaining steps of the proof.

Suppose that we apply the $T^H_{\mu_{k+1}}$ operator $\ell$ times to both sides. Then, due to monotonicity and the fact $T^H_{\mu}(J + \rho\epsilon) = T^H_{\mu}(J) + \rho\epsilon$, for any policy $\mu$, we have the following:

$$T^H_{\mu_{k+1}} T^H_{\mu_{k+2}} \ldots T^H_{\mu_{k+\ell}} J^\mu_k \leq \alpha^{H-1} c_{m,H,\gamma} e + T^{H-1} J^\mu_k.$$

Using a telescoping sum, we get the following inequality:

$$T^H_{\mu_{k+1}} T^H_{\mu_{k+2}} \ldots T^H_{\mu_{k+\ell}} J^\mu_k \leq J^\mu_{k+\ell} e e$$

Taking the limit as $\ell \to \infty$ on both sides, we have the following:

$$J^\mu_{k+1} - J^\star \leq \frac{c_{m,H,\gamma} e}{1 - \alpha} $$

The rest of proof is straightforward. Subtract $J^\star$ from both sides of the previous inequality and using the contraction property of $T$, we get

$$J^\mu_{k+1} - J^\star \leq \frac{c_{m,H,\gamma} e}{1 - \alpha} $$

which is equivalent to

$$\|J^\mu_{k+1} - J^\star\|_\infty \leq \alpha^{H-1} \|J^\mu_k - J^\star\|_\infty e + \frac{c_{m,H,\gamma} e}{1 - \alpha},$$

(14) since $J^\star \leq J^\mu$ for all policies $\mu$. Iterating over $k$, we get Theorem 2.

**Remark 1.** As mentioned earlier, the proof of Theorem 7 is essentially a special case of Theorem 2 with small modifications. The proof of Theorem 7 skips Steps 1 and 2 and instead uses techniques in Lemma 3 (see Appendix A) to obtain an analogous result to Step 3. The rest of the proof (Steps 4-6) follows except with $p_{m,\gamma}$ (defined in the proof of Theorem 2) in place of $p_{m,H,\gamma}$ as defined in Step 3.

### 5 Related Work

In the introduction, we compared our results to those in (Efroni et al. [2019]). Now, we compare our work to other papers in the literature. It is worth noting that (Efroni et al. [2019]) builds on a lot of ideas in (Efroni et al. [2018a]). The role of lookahead and return in improving the performance of RL algorithms has also been studied in a large number

...
of papers including (Moerland, Broekens, and Jonker 2020; Efroni, Ghavamzadeh, and Mannor 2020; Tomar, Efroni, and Ghavamzadeh 2020; Efroni et al. 2018b; Springenberg et al. 2020; Deng et al. 2020). The works of (Baxter, Tridgell, and Weaver 1999; Veness et al. 2009; Lanctot et al. 2014) explore the role of tree search in RL algorithms. However, to the best of our knowledge, the amount of lookahead and return needed as a function of the feature vectors has not been quantified in prior works.

Approximate policy iteration is a well studied topic, see (Bertsekas and Tsitsiklis 1996; Bertsekas 2019; Puterman and Shin 1978; Lesner and Scherrer 2015), for example. However, since their models do not involve a scheme for approximating the value function at each iteration, the role of the depth of lookahead ($H$) and return ($m$) cannot be quantified using their works. It is important to note that our work is different from least squares policy iteration (LSPI), which is a common means of obtaining function approximation parameters from value function estimates. While one of our algorithms used least squares estimates, our work more importantly analyzes the use of lookahead and $m$-step return, which are employed prior to the least squares step of the algorithm.

The works of (Bertsekas 2011) and (Bertsekas 2019) also study a variant of policy iteration wherein a greedy policy is evaluated approximately using feature vectors at each iteration. These papers also provides rates of convergence as well as a bound on the approximation error. However, our main goal is to understand the relations between function approximation and lookahead/return which are not considered in these other works.

6 Conclusion

We show that a minimum threshold for the lookahead and corresponding policy return must be met for algorithms with function approximation based approximate policy iteration to converge. In particular, we show that if one uses function approximation in conjunction with an $m$-step return of an $H$-step lookahead policy without the $T^{H-1}$ factor in the iterates, the $m + H$ must be sufficiently large. Our results have been derived for two different implementations of the function approximation algorithm: by solving the least squares problem or by one step of gradient descent of the least squares objective in each iteration.

As mentioned in the introduction, it would be interesting to extend our work to study neural network based function approximation. Additionally, for problems with a terminal state, it would be interesting to consider cases where the value function of a given policy is estimated using a full rollout which provides an unbiased estimate as in (Tsitsiklis 2002).

7 Acknowledgements

The research presented here was supported in part by a grant from Sandia National Labs and the NSF Grants CCF 1934986, CCF 1704970, ONR Grant N00014-19-1-2566, NSF/USDA Grant AG 2018-67007-28379, and ARO Grant W911NF-19-1-0379.

---

1Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.
References

Baxter, J.; Tridgell, A.; and Weaver, L. 1999. TDLeaf(lambda): Combining Temporal Difference Learning with Game-Tree Search. CoRR, cs.LG/9901010.

Bertsekas, D. 2011. Approximate policy iteration: a survey and some new methods. Journal of Control Theory and Applications, 9: 310–335.

Bertsekas, D. 2021. Lessons from AlphaZero for Optimal, Model Predictive, and Adaptive Control. arXiv:2108.10315.

Bertsekas, D.; and Tsitsiklis, J. 1996. Neuro-dynamic Programming. Athena Scientific. ISBN 9781886529106.

Bertsekas, D. P. 2019. Reinforcement learning and optimal control. Athena Scientific Belmont, MA.

Browne, C.; Powley, E.; Whitehouse, D.; Lucas, S.; Cowling, P.; Rohlfshagen, P.; Tavener, S.; Perez Liebana, D.; Samothrakis, S.; and Colton, S. 2012. A Survey of Monte Carlo Tree Search Methods. IEEE Transactions on Computational Intelligence and AI in Games, 4:1: 1–43.

Cao, Y.; and Gu, Q. 2019. Generalization bounds of stochastic gradient descent for wide and deep neural networks. Advances in Neural Information Processing Systems, 32: 10836–10846.

Deng, H.; Yin, S.; Deng, X.; and Li, S. 2020. Value-based Algorithms Optimization with Discounted Multiple-step Learning Method in Deep Reinforcement Learning. In 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 979–984.

Efroni, Y.; Dalal, G.; Scherrer, B.; and Mannor, S. 2018a. Beyond the One Step Greedy Approach in Reinforcement Learning. CoRR, abs/1802.03654.

Efroni, Y.; Dalal, G.; Scherrer, B.; and Mannor, S. 2018b. Multiple-Step Greedy Policies in Online and Approximate Reinforcement Learning. arXiv:1805.07956.

Efroni, Y.; Dalal, G.; Scherrer, B.; and Mannor, S. 2019. How to Combine Tree-Search Methods in Reinforcement Learning. arXiv:1809.01843.

Efroni, Y.; Ghavamzadeh, M.; and Mannor, S. 2020. Online Planning with Lookahead Policies. Advances in Neural Information Processing Systems, 33.

Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural tangent kernel: Convergence and generalization in neural networks. arXiv preprint arXiv:1806.07572.

Kocsis, L.; and Szepesvári, C. 2006. Bandit Based Monte-Carlo Planning. volume 2006, 282–293. ISBN 978-3-540-45375-8.

Lanctot, M.; Winands, M. H. M.; Pepels, T.; and Sturtevant, N. R. 2014. Monte Carlo Tree Search with Heuristic Evaluations using Implicit Minimax Backups. arXiv:1406.0486.

Lesner, B.; and Scherrer, B. 2015. Non-Stationary Approximate Modified Policy Iteration.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T. P.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous Methods for Deep Reinforcement Learning. CoRR, abs/1602.01783.

Moerland, T. M.; Broekens, J.; and Jonker, C. M. 2020. A Framework for Reinforcement Learning and Planning. arXiv:2006.15009.

Puterman, M.; and Shin, M. C. 1978. Modified Policy Iteration Algorithms for Discounted Markov Decision Problems. Management Science, 24: 1127–1137.

Shah, D.; Xie, Q.; and Xu, Z. 2020. Non-Asymptotic Analysis of Monte Carlo Tree Search. arXiv:1902.05213.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T. P.; Simonyan, K.; and Hassabis, D. 2017a. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. CoRR, abs/1712.01815.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017b. Mastering the game of go without human knowledge. Nature, 550(7676): 354–359.

Springenberg, J. T.; Heess, N.; Mankowitz, D.; Merel, J.; Byravan, A.; Abdolmaleki, A.; Kay, J.; Degrave, J.; Schrittwieser, J.; Tassa, Y.; et al. 2020. Local search for policy iteration in continuous control. arXiv preprint arXiv:2010.05545.

Sutton, R. S.; and Barto, A. G. 2018. Reinforcement Learning: An Introduction. The MIT Press, second edition.

Tomar, M.; Efroni, Y.; and Ghavamzadeh, M. 2020. Multi-step Greedy Reinforcement Learning Algorithms. arXiv:1910.02919.

Tsitsiklis, J. N. 2002. On the convergence of optimistic policy iteration. Journal of Machine Learning Research, 3(Jul): 59–72.

Tsitsiklis, J. N.; and van Roy, B. 1994. Feature-Based Methods For Large Scale Dynamic Programming. In Machine Learning, 59–94.

Veness, J.; Silver, D.; Uther, W.; and Blair, A. 2009. Bootstrapping from Game Tree Search. 1937–1945.
A Proof of Theorem 1

Proof. Consider policies $\mu_k$ and $\mu_{k+1}$, where $\mu_0$ can be taken to be any arbitrary policy. We have the following:

$$T_{\mu_{k+1}} T^{H-1} J^{\mu_k} = T_{\mu_{k+1}} T^{H-1} J^{\mu_k} - T_{\mu_{k+1}} T^{H-1} J_k$$

$$+ T_{\mu_{k+1}} T^{H-1} J_k$$

$$(a) \leq \alpha^H \|J^{\mu_k} - J_k\|_{\infty} e + T_{\mu_{k+1}} T^{H-1} J_k$$

$$\leq \alpha^H \|J^{\mu_k} - J_k\|_{\infty} e + T^{H} J_k + \varepsilon e$$

$$= \alpha^H \|J^{\mu_k} - J_k\|_{\infty} e + T^{H} J^{\mu_k} + \varepsilon e$$

$$(b) \leq 2\alpha^H \|J^{\mu_k} - J_k\|_{\infty} e + T^{H-1} J^{\mu_k} + \varepsilon e$$

where $e$ is the vector of all 1s, (a) and (b) follow from the contraction property of $T^H$ and $T_{\mu_{k+1}}$ and the last inequality follows from standard arguments using the monotonicity properties and the definition of $T$: specifically, note that

$$T J^\mu = \min_{\mu'} T_{\mu'} J^{\mu'} \leq T_{\mu} J^{\mu} = J^\mu,$$

and repeatedly apply $T$ to both sides of the inequality and use monotonicity to obtain $T^\ell J^\mu \leq T^{\ell-1} J^\mu$ for all $\ell$ and all policies $\mu$.

We can further bound $\|J^{\mu_k} - J_k\|$ as follows:

Lemma 3.

$$\|J^{\mu_k} - J_k\| \leq \alpha^{m+H-1} \delta_1 \|J_{k-1} - J^{\mu_{k-1}}\|_{\infty}$$

$$+ \frac{\alpha^{m+H-1}}{1 - \alpha} \delta_1 + \delta_2 + \delta_1 \varepsilon'.$$
Assumptions 1, 2, and 3 hold, we have the following:

\[ \|J^\mu_k - J_k\|_\infty \leq \frac{c_m + \alpha^{m+H-1} \delta_1 + \delta_2 + \delta_1 \varepsilon'}{1 - \alpha^{m+H-1} \delta_1} + \frac{1}{1 - \alpha} + \|J_0\|_\infty, \]  

(18)

where we have used the assumption \( \alpha^{m+H-1} \delta_1 < 1 \) and the fact that \( \|J^\mu_0\|_\infty \leq 1/(1 - \alpha) \) due to Assumption 3.

Putting (17) and (18) together, we have the following:

\[ T_{\mu_{k+1}} T^{H-1} J^\mu_k \leq \left[ 2\alpha H (p_{m,H}) + \varepsilon \right] e + T^{H-1} J^\mu_k. \]  

(19)

The rest of the proof is similar to the arguments in (Bertsekas and Tsitsiklis 1996), (Bertsekas 2019); however, we have to explicitly incorporate the role of lookahead \((H)\) in the remaining steps of the proof.

Suppose that we apply the \( T_{\mu_{k+1}} \) operator \( \ell - 1 \) times to both sides. Then, due to monotonicity and the fact \( T_\mu(J + ce) = T_\mu(J) + \alpha ce \), for any policy \( \mu \), we have the following:

\[ T_{\mu_{k}}^\ell T^{H-1} J^\mu_k \leq \alpha^{\ell-1} c_{m,H} e + T_{\mu_{k+1}}^\ell T^{H-1} J^\mu_k. \]

Using a telescoping sum, we get the following inequality:

\[ T_{\mu_{k+1}}^j T^{H-1} J^\mu_k - T^{H-1} J^\mu_k \leq \sum_{\ell=1}^{j} \alpha^{\ell-1} c_{m,H} e. \]

Taking the limit as \( j \to \infty \) on both sides, we have the following:

\[ J^{\mu_{k+1}} - T^{H-1} J^\mu_k \leq \frac{c_{m,H} e}{1 - \alpha}. \]

Rearranging terms and subtracting \( J^* \) from both sides, we get the following:

\[ J^{\mu_{k+1}} - J^* \leq T^{H-1} J^\mu_k - J^* + \frac{c_{m,H} e}{1 - \alpha} \]

\[ = T^{H-1} J^\mu_k - T^{H-1} J^* + \frac{c_{m,H} e}{1 - \alpha} \]

\[ \leq \alpha^{H-1} \|J^\mu_k - J^*\|_\infty e + \frac{c_{m,H} e}{1 - \alpha}. \]

Since \( J^* \leq J^{\mu_{k+1}} \), we have the following:

\[ \|J^{\mu_{k+1}} - J^*\|_\infty \leq \alpha^{H-1} \|J^\mu_k - J^*\|_\infty + \frac{c_{m,H} e}{1 - \alpha}. \]  

(20)

Iterating over \( k \) and taking limits, we get Theorem 1.

The bound in Equation (20) gives us the following:

\[ \|J^\mu_k - J^*\|_\infty \leq \alpha^{(H-1)k} \|J^\mu_0 - J^*\|_\infty + \sum_{i=0}^{k-1} \alpha^{(H-1)i} \frac{c_{m,H} e}{1 - \alpha}, \]  

(21)

for \( k > 0 \).

Note that the inequality in (21) provides finite-time performance bounds, in addition to the asymptotic result stated in Theorem 1.

\section{A Modified Least Squares Algorithm}

Suppose Step 2 of Algorithm 1 is changed to \( \hat{J}_{\mu_{k+1}}^i = T_{\mu_{k+1}}^m(J_k)(i) + w_{k+1}(i) \) for \( i \in D_k \). Then, it is still possible to get bounds on the performance of the algorithm when \( m \) is sufficiently large. With this modification to the algorithm, when Assumptions 1-2 and 3 hold, we have the following:
Theorem 3. Suppose that $m$ satisfies $m > \log(\delta_1) / \log(1/\alpha)$, where
\[
\delta_1 := \sup_k \|\mathcal{M}_k\|_\infty = \sup_k \|\Phi(D_k^T \Phi D_k)^{-1} \Phi D_k^T P_k\|_\infty,
\]
then
\[
\limsup_{k \to \infty} \|J^{\mu_k} - J^*\|_\infty \leq \frac{\tilde{c}_{m,H}}{(1 - \alpha)(1 - \alpha^{H-1})},
\]
where
\[
\tilde{c}_{m,H} := 2\alpha^H \left( \frac{\alpha^m \delta_2 + \delta_1 \varepsilon'}{1 - \alpha^m} + \frac{1}{1 - \alpha} + \|J_0\|_\infty \right) + \varepsilon
\]
and $\delta_2 := \sup_{k,\mu_k} \|\mathcal{M}_k J^{\mu_k} - J^{\mu_k}\|_\infty$.

The proof of Theorem 3 is as follows:

Proof. Consider policies $\mu_k$ and $\mu_{k+1}$, where $\mu_0$ can be taken to be any arbitrary policy. We have the following:
\[
T_{\mu_{k+1}} T^H J^{\mu_k} = T_{\mu_{k+1}} T^H J^{\mu_k} - T_{\mu_{k+1}} T^H J_k + T_{\mu_{k+1}} T^H J_k,
\]

(a) \[ \leq \alpha^H \|J^{\mu_k} - J_k\|_\infty e + T_{\mu_{k+1}} T^H J_k \]

(b) \[ \leq 2\alpha^H \|J^{\mu_k} - J_k\|_\infty e + T^H J^{\mu_k} + \varepsilon e \]

and repeatedly apply $T$ to both sides of the inequality and use monotonicity to obtain $T^\ell J^{\mu_k} \leq T^{\ell-1} J^{\mu_k}$ for all $\ell$ and all policies $\mu$.

We can further bound $\|J^{\mu_k} - J_k\|$ as follows:
\[
\|J^{\mu_k} - J_k\| = \|\mathcal{M}_k (T_{\mu_k} T_{\mu_{k+1}} J_{k-1} + w_{k}) - J^{\mu_k}\|_\infty
\]

\[ \leq \|\mathcal{M}_k T_{\mu_k} J_{k-1} - J^{\mu_k}\|_\infty + \|\mathcal{M}_k w_{k}\|_\infty \]

\[ \leq \|\mathcal{M}_k T_{\mu_k} J_{k-1} - J^{\mu_k}\|_\infty + \|\mathcal{M}_k\|_\infty \|w_{k}\|_\infty \]

\[ \leq \|\mathcal{M}_k T_{\mu_k} J_{k-1} - J^{\mu_k}\|_\infty + \delta_1 \varepsilon' \]

\[ = \|\mathcal{M}_k T_{\mu_k} J_{k-1} - \mathcal{M}_k J^{\mu_k} + \mathcal{M}_k J^{\mu_k} - J^{\mu_k}\|_\infty + \delta_1 \varepsilon' \]

\[ \leq \|\mathcal{M}_k T_{\mu_k} J_{k-1} - \mathcal{M}_k J^{\mu_k}\|_\infty + \|\mathcal{M}_k J^{\mu_k} - J^{\mu_k}\|_\infty + \delta_1 \varepsilon' \]

\[ \leq \delta_1 \|T_{\mu_k} J_{k-1} - J^{\mu_k}\|_\infty + \|\mathcal{M}_k J^{\mu_k} - J^{\mu_k}\|_\infty + \delta_1 \varepsilon' \]

\[ \leq \alpha^m \delta_1 \|J_{k-1} - J^{\mu_k}\|_\infty + \sup_{k,\mu_k} \|\mathcal{M}_k J^{\mu_k} - J^{\mu_k}\|_\infty + \delta_1 \varepsilon' \]

\[ \leq \alpha^m \delta_1 \|J_{k-1} - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon' \]

\[ \leq \alpha^m \delta_1 \|J_{k-1} - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon' \]

\[ \leq \alpha^m \delta_1 \|J_{k-1} - J^{\mu_k}\|_\infty + \alpha^m \delta_1 \|J^{\mu_{k-1}} - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon' \]

\[ \leq \alpha^m \delta_1 \|J_{k-1} - J^{\mu_{k-1}}\|_\infty + \frac{\alpha^m \delta_1}{1 - \alpha} + \delta_2 + \delta_1 \varepsilon' \]

Iterating over $k$ we get that for all $k$,
\[
\|J^{\mu_k} - J_k\|_\infty \leq \frac{\alpha^m \delta_1 + \delta_2 + \delta_1 \varepsilon'}{1 - \alpha^m} + \frac{1}{1 - \alpha} + \|J_0\|_\infty, \tag{23}
\]
where we have used the assumption \( \alpha^m \delta_1 < 1 \) and the fact that \( \| J^{\mu_0} \|_\infty \leq 1/(1 - \alpha) \) due to Assumption 3.

Putting (22) and (23) together, we have the following:

\[
T_{\mu_{k+1}} T^{H-1} J^{\mu_k} \leq \left[ 2\alpha^H p_m + \varepsilon \right] e + T^{H-1} J^{\mu_k}.
\]

The rest of the proof follows from the proof of Theorem 1 with \( \tilde{c}_{m,H} \) instead of \( c_{m,H} \) and we get Theorem 3.

Analogously to the inequality (21) in Appendix A, the proof of Theorem 3 gives us the following:

\[
\| J^{\mu_k} - J^* \|_\infty \leq \alpha^{(H-1)k} \| J^{\mu_0} - J^* \|_\infty + \sum_{i=0}^{k-1} \alpha^{(H-1)i} \tilde{c}_{m,H} 1^{-\alpha},
\]

for \( k > 0 \).

Note that the inequality (24) provides finite-time performance bounds for the modified least squares algorithm, in addition to the asymptotic result stated in Theorem 3.

C Proof of Proposition 1

In the following proposition, we present a bound on the difference between \( J_k \) and \( J^* \).

**Proposition 1.** When \( \alpha^m H - 1 \delta_1 < 1 \), \( \lim \sup_{k \to \infty} \| J_k - J^* \|_\infty \leq \frac{\tilde{c}_{m,H}}{1 - \alpha} \),

where \( q_{m,H} := \delta_1 \alpha^m H - 1 + (1 + \delta_1 \alpha^m) \alpha H - 1 \) and \( r_{m,H} := (1 + \delta_1 \alpha^m) \left( \alpha H - 1 p_m H + \frac{c_{m,H}}{1 - \alpha} \right) + \delta_2 + \delta_1 \varepsilon' \), where \( c_{m,H} \) is defined in (19) and \( p_m,H \) is defined in (18). The proof is as follows.

**Proof.**

\[
\| J_{k+1} - J^* \|_\infty = \| J_{k+1} - J^{\mu_{k+1}} + J^{\mu_{k+1}} - J^* \|_\infty
\]

\[
\leq \| J_{k+1} - J^{\mu_{k+1}} \|_\infty + \| J^{\mu_{k+1}} - J^* \|_\infty
\]

\[
\leq \| M_{k+1} T^{H-1} J_k - J^{\mu_{k+1}} \|_\infty + \| J^{\mu_{k+1}} - J^* \|_\infty + \delta_1 \varepsilon'
\]

\[
= \| M_{k+1} T A_{k+1} \|_\infty \| T^{H-1} J_k - M_{k+1} J^{\mu_{k+1}} \|_\infty + \| M_{k+1} J^{\mu_{k+1}} - J^{\mu_{k+1}} \|_\infty + \| J^{\mu_{k+1}} - J^* \|_\infty + \delta_1 \varepsilon'
\]

\[
\leq \| M_{k+1} \|_\infty \| T^{H-1} J_k - M_{k+1} J^{\mu_{k+1}} \|_\infty + \| M_{k+1} J^{\mu_{k+1}} - J^{\mu_{k+1}} \|_\infty + \| J^{\mu_{k+1}} - J^* \|_\infty + \delta_1 \varepsilon'
\]

\[
\leq \delta_1 \alpha^m \| T^{H-1} J_k - J^{\mu_{k+1}} \|_\infty + \delta_2 + \| J^{\mu_{k+1}} - J^* \|_\infty + \delta_1 \varepsilon'
\]

\[
= \delta_1 \alpha^m \| T^{H-1} J_k - J^{\mu_{k+1}} \|_\infty + \delta_2 + \| J^{\mu_{k+1}} - J^* \|_\infty + \delta_1 \varepsilon'
\]

\[
\leq \delta_1 \alpha^m \| T^{H-1} J_k - J^* \|_\infty + \delta_2 + \| J^{\mu_{k+1}} - J^* \|_\infty + \delta_1 \varepsilon'
\]

\[
\leq \delta_1 \alpha^m \| T^{H-1} J_k - J^* \|_\infty + \delta_2 + \| J^{\mu_{k+1}} - J^* \|_\infty + \delta_1 \varepsilon'
\]

\[
\leq \delta_1 \alpha^m \| J_k - J^* \|_\infty + (1 + \delta_1 \alpha^m) \| J^{\mu_{k+1}} - J^* \|_\infty + \delta_2 + \delta_1 \varepsilon'
\]

\[
\leq \delta_1 \alpha^m \left( \alpha H - 1 \| J^{\mu_k} - J^* \|_\infty + \frac{c_{m,H}}{1 - \alpha} \right) + \delta_2 + \delta_1 \varepsilon'
\]

\[
\leq \delta_1 \alpha^m \left( \alpha H - 1 \| J^{\mu_k} - J^* \|_\infty + \frac{c_{m,H}}{1 - \alpha} \right) + \delta_2 + \delta_1 \varepsilon'
\]

\[
= \left( \delta_1 \alpha^m H - 1 + (1 + \delta_1 \alpha^m) \alpha H - 1 \right) \| J_k - J^* \|_\infty + (1 + \delta_1 \alpha^m) \left( \alpha H - 1 \| J^{\mu_k} - J^* \|_\infty + \frac{c_{m,H}}{1 - \alpha} \right) + \delta_2 + \delta_1 \varepsilon'
\]

\[
\leq \delta_1 \alpha^m H - 1 + (1 + \delta_1 \alpha^m) \alpha H - 1 \| J_k - J^* \|_\infty + (1 + \delta_1 \alpha^m) \left( \alpha H - 1 \| J^{\mu_k} - J^* \|_\infty + \frac{c_{m,H}}{1 - \alpha} \right) + \delta_2 + \delta_1 \varepsilon'
\]

\[
= \delta_1 \alpha^m H - 1 \| J_k - J^* \|_\infty + r_{m,H}.
\]
where \( q_{m,H} := \delta_1 \alpha^{m+H-1} + (1 + \delta_1 \alpha^m) \alpha^{H-1} \) and \( r_{m,H} := (1 + \delta_1 \alpha^m) \left( \alpha^{H-1} p_{m,H} + \frac{c_{m,H}}{1-\alpha} \right) + \delta_2 + \delta_1 \varepsilon'. \) Note that \( p_{m,H} \) is defined in (18) and \( c_{m,H} \) is defined in (19). Additionally, \((c)\) follows from (20) and \((d)\) follows from (18). Iterating over \( k \), we get Proposition I.

We obtain the following inequality in much a similar way to inequalities (21) and (36) in the proofs of Theorem I and Theorem III respectively:

\[
\| J_k - J^* \|_{\infty} \leq q_{m,H}^k \| J_0 - J^* \|_{\infty} + \sum_{i=0}^{k-1} q_{m,H}^i r_{m,H},
\]

for \( k > 0 \). Note that the inequality (25) provides finite-time performance bounds, in addition to the asymptotic result stated in Proposition I.

\section*{D Proof of Lemma I}

\textit{Proof.}

\[
\frac{1}{1-\alpha} \geq \| J^{\mu_k} - J^{\mu_{k+1}} \|_{\infty}
\]

\[
= \| J^{\mu_k} - \Phi \hat{\mu}_k - \left( J^{\mu_{k-1}} - \Phi \hat{\mu}_{k-1} + \Phi \theta^{\mu_{k-1}} - \Phi \theta^{\mu_k} \right) \|_{\infty}
\]

\[
\geq \| J^{\mu_{k-1}} - \Phi \theta^{\mu_{k-1}} + \Phi \theta^{\mu_{k-1}} - \Phi \theta^{\mu_k} \|_{\infty} - \delta'_2
\]

\[
\geq \frac{1}{\sigma_{\min, \Phi}} \| \theta^{\mu_{k-1}} - \theta^{\mu_k} \|_{\infty} - 2\delta'_2
\]

\[
\Rightarrow \sigma_{\min, \Phi} \frac{1}{1-\alpha} + 2\sigma_{\min, \Phi} \delta'_2 \geq \| \theta^{\mu_{k-1}} - \theta^{\mu_k} \|_{\infty},
\]

where \( \sigma_{\min, \Phi} \) is the smallest singular value in the singular value decomposition of \( \Phi \).

\section*{E Proof of Lemma II}

\textit{Proof.} From assumption I we have that \( \Phi_{D_k} \) is of rank \( d \). Thus, we have the following:

\[
\hat{\mu}_k - \tilde{\mu}_k = \left( \Phi_{D_k}^T \Phi_{D_k} \right)^{-1} \Phi_{D_k}^T \left( \mathcal{P}_k J^{\mu_k} - \mathcal{P}_k (T_{\mu_k}^{m,H-1}J_{k-1} + w_k) \right)
\]

\[
= \left( \Phi_{D_k}^T \Phi_{D_k} \right)^{-1} \Phi_{D_k}^T \mathcal{P}_k \left( J^{\mu_k} - \left( T_{\mu_k}^{m,H-1} \Phi \theta_{k-1} + w_k \right) \right)
\]

\[
\Rightarrow \| \hat{\mu}_k - \tilde{\mu}_k \|_{\infty} \leq \left\| \left( \Phi_{D_k}^T \Phi_{D_k} \right)^{-1} \Phi_{D_k}^T \| \right\|_{\infty} \| J^{\mu_k} - \left( T_{\mu_k}^{m,H-1} \Phi \theta_{k-1} + w_k \right) \|_{\infty}
\]

\[
\leq \left\| \left( \Phi_{D_k}^T \Phi_{D_k} \right)^{-1} \Phi_{D_k}^T \| \right\|_{\infty} \| J^{\mu_k} - T_{\mu_k}^{m,H-1} \Phi \theta_{k-1} \|_{\infty} + \frac{\delta'_1 \varepsilon'}{\| \Phi \|_{\infty}}
\]

\[
\leq \alpha^{m} \sup_k \left\| \left( \Phi_{D_k}^T \Phi_{D_k} \right)^{-1} \Phi_{D_k}^T \right\|_{\infty} \| J^{\mu_k} - T^{H-1} \Phi \theta_{k-1} \|_{\infty} + \frac{\delta'_1 \varepsilon'}{\| \Phi \|_{\infty}}
\]

(26)
We now obtain a bound for \( \| J^{\mu_k} - T^{H-1} \Phi \theta_{k-1} \|_{\max} \) as follows:

\[
\| J^{\mu_k} - T^{H-1} \Phi \theta_{k-1} \|_{\max} = \| J^{\mu_k} - J^* + J^* - T^{H-1} \Phi \theta_{k-1} \|_{\max} \\
\leq \| J^{\mu_k} - J^* \|_{\max} + \| J^* - T^{H-1} \Phi \theta_{k-1} \|_{\max} \\
\leq \| J^{\mu_k} - J^* \|_{\max} + \alpha H^{-1} \| J^* - \Phi \theta_{k-1} \|_{\max} \\
\leq \frac{1}{1 - \alpha} + \alpha H^{-1} \| J^* - \Phi \theta_{k-1} \|_{\max} \\
\leq \frac{1}{1 - \alpha} + \alpha H^{-1} \| J^* - J^{\mu_k-1} + J^{\mu_k-1} - \Phi \theta_{k-1} \|_{\max} \\
\leq \frac{1}{1 - \alpha} + \alpha H^{-1} \left( \| J^* - J^{\mu_k-1} \|_{\max} + \| J^{\mu_k-1} - \Phi \theta_{k-1} \|_{\max} \right) \\
\leq \frac{1}{1 - \alpha} + \alpha H^{-1} \| J^{\mu_k-1} - \Phi \theta_{k-1} \|_{\max} \\
= \frac{1}{1 - \alpha} + \alpha H^{-1} \| J^{\mu_k-1} - \Phi \theta_{k-1} + \Phi \theta_{k-1} - \Phi \theta_{k-1} \|_{\max} \\
= \frac{1}{1 - \alpha} + \alpha H^{-1} \| J^{\mu_k-1} - \Phi \theta_{k-1} \|_{\max} + \alpha H^{-1} \| \Phi \theta_{k-1} - \Phi \theta_{k-1} \|_{\max} \\
\leq \frac{1}{1 - \alpha} + \alpha H^{-1} \| J^{\mu_k-1} \|_{\max} + \| \theta_{k-1} \|_{\max}.
\]

Putting (26) and (27) together gives the following:

\[
\| \theta^{\mu_k} - \hat{\theta}^{\mu_k} \|_{\max} \leq \frac{\delta_1 \epsilon'}{\| \Phi \|_{\max}} + \alpha \frac{\delta_2}{\| \Phi \|_{\max}} \left( \frac{1 + \alpha H^{-1}}{1 - \alpha} + \alpha H^{-1} \delta_2 \right) + \alpha^{m + H^{-1}} \| \theta^{\mu_k-1} - \theta_{k-1} \|_{\max}.
\]

\( \square \)

**F TD-Learning Approximation**

We consider the variant of approximate policy iteration studied in (Efroni et al. 2019). We define the following operator with parameter \( \lambda \in (0, 1) \):

\[
T^{\mu_k}_\lambda J = (1 - \lambda) \sum_{j=0}^{\infty} \lambda^j (T^{\mu}_\lambda)^{j+1} J \\
= J + (1 - \gamma \lambda P^{\mu_k-1}) (T^{\mu_k}_\lambda J - J),
\]

where \( P^{\mu} \) is the state transition matrix corresponding policy \( \mu \). This operator is an estimator of the TD \( - \lambda \) operator in (Sutton and Barto 2018).

The full return \( (m = \infty) \) is often desired in practice but is impossible to obtain for some MDPs. It is possible to instead obtain an estimate of a full return for any policy \( \mu_k \) with the operator in equation (28) \( T^{\mu_k}_\lambda \) parameterized by \( \lambda \in (0, 1) \). Using the \( T^{\mu_k}_\lambda \) operator to obtain an estimate for \( J^{\mu_k} \), our modified algorithm is given in Algorithm 3.

As was the case with our main algorithm, we make Assumption 1, Assumption 2, and Assumption 3. Using Assumption 1, we can succinctly write our iterates as follows:

\[
J_{k+1} = M_{k+1} (T^{\mu_k+i}_\lambda T^{H-1} J_k + w_{k+1}).
\]

We will now state a theorem that characterizes the role of \( \lambda \) in the convergence of our algorithm:

**Theorem 4.** When \( \delta_1 c_{\lambda} H^{-1} < 1 \), where \( \delta_1 \) is defined in Theorem 1 and \( c_{\lambda} := \frac{1 - \lambda}{\lambda (1 - \lambda \alpha)} \),

\[
\limsup_{k \to \infty} \| J^{\mu_k} - J^* \|_{\max} \leq \frac{c_{m,H,\lambda}}{(1 - \alpha)(1 - \alpha H^{-1})},
\]

where \( c_{m,H,\lambda} := 2\alpha H \left( \frac{\delta_1 c_{\lambda} H^{-1} + \delta_1 + \delta_2 \epsilon'}{1 - \delta_1 c_{\lambda} H^{-1}} \right) + \frac{1}{1 - \alpha} + \| J_0 \|_{\max} \), and \( \delta_2 \) is defined in Theorem 1.

The proof of Theorem 4 is as follows:
which we evaluate the current policy at iteration \( k \).

**Proof.** Consider the following sequences:

\[
\begin{align*}
V_k^i &:= (1 + \lambda + \ldots + \lambda^i)^{-1} \sum_{j=0}^{i} \lambda^j (T^{\mu_k})^{j+1} (T^{H-1} J_{k-1} - J^{\mu_k}) \\
z_k^i &:= V_k^i - J^{\mu_k}.
\end{align*}
\]

First, we get bounds on \( V_k^i \). We have the following:

\[
\begin{align*}
J^{\mu_k} - \|T^{H-1} J_{k-1} - J^{\mu_k}\|_{\infty} &\leq T^{H-1} J_{k-1} - J^{\mu_k} \leq J^{\mu_k} + \|T^{H-1} J_{k-1} - J^{\mu_k}\|_{\infty} e \\
\Rightarrow (1 + \lambda + \ldots + \lambda^i)^{-1} \sum_{j=0}^{i} \lambda^j (T^{\mu_k})^{j+1} (J^{\mu_k} - \|T^{H-1} J_{k-1} - J^{\mu_k}\|_{\infty} e) - J^{\mu_k} \\
&\leq z_k^i \leq (1 + \lambda + \ldots + \lambda^i)^{-1} \sum_{j=0}^{i} \lambda^j (T^{\mu_k})^{j+1} (J^{\mu_k} + \|T^{H-1} J_{k-1} - J^{\mu_k}\|_{\infty} e) - J^{\mu_k} \\
\Rightarrow (1 + \lambda + \ldots + \lambda^i)^{-1} \sum_{j=0}^{i} \lambda^j (T^{\mu_k})^{j+1} (J^{\mu_k} - \varepsilon \lambda e) - J^{\mu_k} \leq z_k^i \\
&\leq (1 + \lambda + \ldots + \lambda^i)^{-1} \sum_{j=0}^{i} \lambda^j (T^{\mu_k})^{j+1} (J^{\mu_k} + \varepsilon \lambda e) - J^{\mu_k} \\
\Rightarrow -(1 + \lambda + \ldots + \lambda^i)^{-1} \sum_{j=0}^{i} \lambda^j \alpha^{j+1} \varepsilon \lambda e \leq z_k^i \leq (1 + \lambda + \ldots + \lambda^i)^{-1} \sum_{j=0}^{i} \lambda^j \alpha^{j+1} \varepsilon \lambda e \\
\Rightarrow \|z_k\|_{\infty} \leq (1 + \lambda + \ldots + \lambda^i)^{-1} \sum_{j=0}^{i} \lambda^j \alpha^{j+1} \varepsilon \lambda.
\end{align*}
\]
Taking limits since norm is a continuous function, we have the following using a “continuity” argument:

\[
\|J_k - J^{\mu_k}\|_\infty = \|\mathcal{M}_k(T^{\mu_k}_\lambda T^{H-1} k + w_k) - J^{\mu_k}\|_\infty + \delta_1 \varepsilon'
\]

\[
\leq \|\mathcal{M}_k T^{\mu_k}_\lambda T^{H-1} k - J^{\mu_k}\|_\infty + \delta_1 \varepsilon'
\]

\[
\leq \|\mathcal{M}_k T^{\mu_k}_\lambda T^{H-1} k - J^{\mu_k}\|_\infty + \|\mathcal{M}_k J^{\mu_k} - J^{\mu_k}\|_\infty + \delta_1 \varepsilon'
\]

\[
\leq \|\mathcal{M}_k\|_\infty \|T^{\mu_k}_\lambda T^{H-1} k - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon'
\]

\[
\leq \delta_1 \|T^{\mu_k}_\lambda T^{H-1} k - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon'
\]

\[
= \delta_1 \lim_i \|V_i^k - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon'
\]

\[
= \delta_1 \lim_i \|z_i^k\|_\infty + \delta_2 + \delta_1 \varepsilon'
\]

\[
\leq \delta_1 \frac{\varepsilon_\lambda (1 - \lambda)}{\lambda (1 - \lambda \alpha)} + \delta_2 + \delta_1 \varepsilon'
\]

\[
= \delta_1 \frac{\|T^{H-1} k - J^{\mu_k}\|_\infty (1 - \lambda)}{\lambda (1 - \lambda \alpha)} + \delta_2 + \delta_1 \varepsilon'.
\]

We now have the following:

\[
\|J_k - J^{\mu_k}\| \leq \delta_1 \|T^{H-1} k - J^{\mu_k}\|_\infty \frac{1 - \lambda}{\lambda (1 - \lambda \alpha)} + \delta_2 + \delta_1 \varepsilon'.
\]  (30)

\[
= \delta_1 c_\lambda \|T^{H-1} k - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon'.
\]  (31)

\[
= \delta_1 c_\lambda \|T^{H-1} k - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon'.
\]

\[
\leq \delta_1 c_\lambda \|T^{H-1} k - J^{\mu_k}\|_\infty + \delta_1 c_\lambda \|J^* - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon'
\]

\[
= \delta_1 c_\lambda \|T^{H-1} k - J^{\mu_k}\|_\infty + \frac{\delta_1 c_\lambda}{1 - \alpha} + \delta_2 + \delta_1 \varepsilon'
\]

\[
= \delta_1 c_\lambda \|T^{H-1} k - J^{\mu_k}\|_\infty + \frac{\delta_1 c_\lambda \alpha^{H-1}}{1 - \alpha} + \delta_2 + \delta_1 \varepsilon'
\]

\[
\leq \delta_1 c_\lambda \|T^{H-1} k - J^{\mu_k}\|_\infty + \delta_1 c_\lambda \left(\frac{1 + \alpha^{H-1}}{1 - \alpha}\right) + \delta_2 + \delta_1 \varepsilon'.
\]

When we iterate over \(k\), we get the following for all \(k\):

\[
\|J^{\mu_k} - J_k\|_\infty \leq \frac{\delta_1 c_\lambda \left(\frac{1 + \alpha^{H-1}}{1 - \alpha}\right) + \delta_2 + \delta_1 \varepsilon'}{1 - \delta_1 c_\lambda \alpha^{H-1}} + \frac{1}{1 - \alpha} + \|J_0\|_\infty
\]

when \(\delta_1 c_\lambda \alpha^{H-1} < 1\).

The rest of the proof follows from the proof of Theorem[1] with \(p_{m,H,\lambda}\) instead of \(p_{m,H}\) and we get Theorem[4] \(\Box\)

The corresponding finite-time bounds are as follows:

\[
\|J^{\mu_k} - J^*\|_\infty \leq \alpha^{(H-1)k} \|J^{\mu_0} - J^*\|_\infty + \sum_{i=0}^{k-1} \alpha^{(H-1)i} \frac{\varepsilon_{m,H,\lambda}}{1 - \alpha},
\]  (32)

for \(k > 0\).

\section{TD-Learning Approximation - A Special Case}

When the \(\lambda\) in our TD-learning algorithm is very small, we may obtain better bounds using an alternative proof technique. Note that in this case, \(T^{\mu_{k+1}}_\lambda\) can be seen as an approximation to the \(T^{\mu_{k+1}}\) operator. We have the following result that is tailored towards small \(\lambda\):
Proposition 2. When \( z_\lambda \alpha^{H-1} < 1 \), where \( z_\lambda := (\delta_1 \delta_3 \alpha + \delta_1 \delta_3 + \alpha), \delta_3 := \|(I - \gamma \lambda P^{\mu_k})^{-1} - I\|_\infty \), and \( \delta_1, \delta_2 \) are defined in Theorem[2]

\[
\limsup_{k \to \infty} \|J^{\mu_k} - J^*\|_\infty \leq \frac{\tilde{c}_{m,H,\lambda}}{(1 - \alpha)(1 - \alpha^{H-1})},
\]

where \( \tilde{c}_{m,H,\lambda} := 2\alpha H \left( \frac{z_\lambda}{1 - z_\lambda \alpha^{H-1}} \right)^{-1} + \frac{1}{1 - \alpha} + \|J_0\|_\infty + \varepsilon. \)

The proof of Proposition[2] is as follows:

Proof. Note that our iterates can be re-written as follows:

\[
J_{k+1} = M_{k+1} \left( T^{H-1}J_k + (I - \gamma \lambda P^{\mu_{k+1}})^{-1}(T^{\mu_{k+1}}T^{H-1}J_k - T^{H-1}J_k) + w_k \right).
\]

We have the following bounds:

\[
\|J^{\mu_k} - J_k\|_\infty = \|M_k \left( T^{H-1}J_{k-1} + (I - \gamma \lambda P^{\mu_k})^{-1}(T^{\mu_k}T^{H-1}J_{k-1} - T^{H-1}J_{k-1}) + w_k \right) - J^{\mu_k}\|_\infty
\]

\[
\leq \|M_k \left( T^{H-1}J_{k-1} + (I - \gamma \lambda P^{\mu_k})^{-1}(T^{\mu_k}T^{H-1}J_{k-1} - T^{H-1}J_{k-1}) \right) - J^{\mu_k}\|_\infty + \delta_1 \varepsilon'
\]

\[
= \|M_k \left( T^{H-1}J_{k-1} + (I - \gamma \lambda P^{\mu_k})^{-1}(T^{\mu_k}T^{H-1}J_{k-1} - T^{H-1}J_{k-1}) \right) - M_kJ^{\mu_k} + M_kJ^{\mu_k} - J^{\mu_k}\|_\infty + \delta_1 \varepsilon'
\]

\[
\leq \|M_k \left( T^{H-1}J_{k-1} + (I - \gamma \lambda P^{\mu_k})^{-1}(T^{\mu_k}T^{H-1}J_{k-1} - T^{H-1}J_{k-1}) \right) - M_kJ^{\mu_k}\|_\infty + \|M_kJ^{\mu_k} - J^{\mu_k}\|_\infty + \delta_1 \varepsilon'
\]

\[
\leq \|M_k \left( T^{H-1}J_{k-1} + (I - \gamma \lambda P^{\mu_k})^{-1}(T^{\mu_k}T^{H-1}J_{k-1} - T^{H-1}J_{k-1}) \right) - M_kJ^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon'
\]

\[
\leq \|M_k \left( T^{H-1}J_{k-1} + (I - \gamma \lambda P^{\mu_k})^{-1}(T^{\mu_k}T^{H-1}J_{k-1} - T^{H-1}J_{k-1}) \right) - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon'
\]

\[
\leq \delta_1 \left( (I - \gamma \lambda P^{\mu_k})^{-1}(T^{\mu_k}T^{H-1}J_{k-1} - T^{H-1}J_{k-1}) \right) - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon'
\]

\[
= \delta_1 \left( (I - \gamma \lambda P^{\mu_k})^{-1}(T^{\mu_k}T^{H-1}J_{k-1} - T^{H-1}J_{k-1}) \right) - J^{\mu_k}\|_\infty + \delta_2 + \delta_1 \varepsilon'
\]

\[
\vdots
\]

(33)
when drawn uniformly from $\mathcal{X}$. In each experiment, a goal state is chosen uniformly at random to have a reward of 1, while each other state has a fixed reward $m_i > 0$.

In this appendix, we test our algorithms on a grid world problem, using the same grid world problem as in (Efroni et al. 2019).

In order to perform linear function approximation, we prescribe a feature vector for each state. In this section, we focus on three particular choices:

1. Random feature vectors: each entry of the matrix $\Phi$ is an independent $\mathcal{N}(0, 1)$ random variable
2. Designed feature vectors: the feature vector for a state whose coordinates are $(x, y)$ is $[x, y, d, 1]^T$, where $d$ is the number of steps required to reach the goal from state $(x, y)$.
3. Indicator vectors: the feature vector for each state $i$ is a $N^2$-dimensional indicator vector where only the $i$-th entry is nonzero.

The random feature vectors represent features which are not necessarily representative of their states. Empirically, although our algorithm will still converge, it will require a large amount of lookahead. By contrast, the designed feature vectors are better able to represent the value function with fewer features, and less lookahead is required. We test Algorithm 1 in each of our experiments, using a starting state of $J_0 = \theta_0 = 0$. All plots in this section graph an average over 20 trials, where each trial has a fixed random choice of $D_k$, the set of states used for policy evaluation. Error bars show the standard deviation of the mean. The code used to produce these graphs is included in the supplementary material.

**H Additional Numerical Experiments**

In this appendix, we test our algorithms on a grid world problem, using the same grid world problem as in (Efroni et al. 2019).

For our simulations, we assume a deterministic grid world problem played on an $N \times N$ grid. The states are the squares of the grid and the actions are {‘up’, ‘down’, ‘right’, ‘left’, and ‘stay’}, which move the agent in the prescribed direction, if possible. In each experiment, a goal state is chosen uniformly at random to have a reward of 1, while each other state has a fixed reward drawn uniformly from $[-0.1, 0.1]$. Unless otherwise mentioned, for the duration of this section, $N = 25$ and $\alpha = 0.9$.

In order to perform linear function approximation, we prescribe a feature vector for each state. In this section, we focus on three particular choices:

1. Random feature vectors: each entry of the matrix $\Phi$ is an independent $\mathcal{N}(0, 1)$ random variable
2. Designed feature vectors: the feature vector for a state whose coordinates are $(x, y)$ is $[x, y, d, 1]^T$, where $d$ is the number of steps required to reach the goal from state $(x, y)$.
3. Indicator vectors: the feature vector for each state $i$ is a $N^2$-dimensional indicator vector where only the $i$-th entry is nonzero.

The random feature vectors represent features which are not necessarily representative of their states. Empirically, although our algorithm will still converge, it will require a large amount of lookahead. By contrast, the designed feature vectors are better able to represent the value function with fewer features, and less lookahead is required.

We test Algorithm 1 in each of our experiments, using a starting state of $J_0 = \theta_0 = 0$. All plots in this section graph an average over 20 trials, where each trial has a fixed random choice of $D_k$, the set of states used for policy evaluation. Error bars show the standard deviation of the mean. The code used to produce these graphs is included in the supplementary material.

### H.1 The effect of $m$ and $H$ on convergence

In Figure 2 we showed how $H$ and $m$ affect convergence of the iterates $J_k$ to $J^*$.

When $m$ and $H$ are small, the value of $J_k$ sometimes diverges. If the value diverges for even one trial, then the average over trials of $\|J_k - J^*\|_\infty$ also increases exponentially with $k$. However, if the average converges for all trials, then the plot is relatively flat. The $m$ or $H$ required for convergence depends on the parameter $\delta_1$ defined in Theorem 1. Over 20 trials, the average value of $\delta_1$ for each of our choices of feature vectors are $30.22, 16.29, \text{and } 1.0$, respectively. As showed through a counter-example in the main body of the paper,
Figure 3: We plot the probability that $\|J_k - J^*\|_\infty$ diverges as a function of $H$ and $m$. For the first plot, $m = 3$ and for the second plot, $H = 3$. In both cases, the algorithm never diverges once $H + m$ is large enough, though a smaller amount of lookahead or $m$-step return are needed for the designed feature vectors.

in general, one needs $m + H - 1 > \log(\delta_1)/\log(1/\alpha)$ for convergence. However, in specific examples, it is possible for convergence to for smaller values of $m + H$. For example, in our grid word model, $\frac{\log(16.29)}{\log(1/0.9)} \approx 26.5$, but we will observe that such a large amount of $m + H$ is not required for convergence.

In Figure 2 it is difficult to see how $H$ and $m$ affect the probability of divergence, as a function of the representative states chosen to be sampled. Therefore, we introduce Figure 3. These plots show the proportion of trials in which the distance $\|J_k - J^*\|_\infty$ exceeded $10^5$ after 30 iterations of our algorithm. As expected, the algorithm never diverges for indicator vectors, as our algorithm is then equivalent to the tabular setting. The designed feature vectors clearly require a much smaller amount of lookahead or $m$-step return, well below the amount predicted by the average $\delta_1$ of 16.29. However, no matter the choice of feature vectors, we will eventually prevent our algorithm from diverging with a large enough value of $H + m$.

H.2 Convergence to the optimal policy

In Theorem 1, we show that as $H$ increases, we converge to a policy $\mu_k$ that is closer to the optimal policy. In this section, we experimentally investigate the role of $m$ and $H$ on the final value of $\|J^{\mu_k} - J^*\|_\infty$. The results can be found in Figure 4. As predicted by theory, we do get closer to the optimal policy as $H$ increases. However, increasing $m$ does not help past a certain point, which is also consistent with the theory. Indeed, although $\mu_k$ is approaching the optimal policy $\mu^*$ as $H$ increases, the iterates $J_k$ are not converging to $J^*$ due to error induced by function approximation. Increasing $m$ improves the policy evaluation, but cannot correct for this inherent error from approximating the value function.
Figure 4: We plot the final value of $\|J^{\mu_k} - J^*\|_\infty$ after 30 iterations. For the first plot, $m = 3$ and for the second plot, $H = 3$. As $H$ increases, the final policy improves. With large enough $H$, we obtain the optimal policy. However, past a certain point, increasing $m$ is not helpful for finding a better policy.