2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

On the Convergence of Natural Policy Gradient and Mirror Descent-Like Policy Methods for Average-Reward MDPs

Yashaswini Murthy and R. Srikant

***Abstract*— It is now well known that Natural Policy Gradient (NPG) globally converges for discounted-reward MDPs in the tabular setting, with perfect value function estimates. However, the result cannot be directly used to obtain a corresponding convergence result for average-reward MDPs by letting the discount factor tend to one. In this paper, we prove that NPG also converges for average-reward MDPs in which each policy leads to an irreducible Markov chain. Since NPG can also be interpreted as a mirror descent based policy method, we then discuss extensions to non-tabular settings for mirror descent-based methods.**

I. Introduction

Popular algorithms to solve for optimal policies in the context of Markov decision processes (MDPs) include dynamic programming based algorithms such as value iteration, policy iteration, and modified policy iteration [1], [2]. While these algorithms require knowledge of the underlying MDP, reinforcement learning algorithms such as policy gradient algorithms can be implemented without the knowledge of an underlying model [3]. A variant of policy gradient methods, called Natural Policy Gradient (NPG) [4], [5] utilizes the Fisher information associated with a policy to condition the gradient. Such a conditioning is known to have good properties, including global convergence when the gradient can be calculated exactly or good performance bounds otherwise [6]. In the tabular setting, NPG has been shown to be closely related to mirror descent [7]–[10]. Mirror descent is a generalization of gradient descent, which uses Bregman divergences as its distance metric and has been studied in various contexts [7], [11]–[16]. In the rest of the paper, we study one specific mirror-descent based method (MDM) which coincides with MDM in the tabular setting.

In the discounted-reward tabular setting, when the value function can be computed exactly, MDM has been shown to converge to the optimal policy with a sublinear rate of $O(\frac{1}{T})$ [6], [17] that is, the MDM update yields the following error bound

$$V_\eta^{\pi^*} - V_\eta^{\pi_T} \leq O\left(\frac{1}{T(1-\alpha)^3}\right),$$

where η is the initial distribution over the state space, π^* is the optimal policy (in fact it can be any arbitrary policy; it is possible to show that the policies obtained through MDM perform better than any arbitrarily chosen policy), V_η^π is the discounted reward associated with policy π. It is well known that the average reward associated with a policy π can be expressed in terms of its discounted reward counterpart as $J^\pi = \lim_{\alpha\to 1}(1-\alpha)V_\eta^\pi(s)$, independent of the state s and initial distribution η [2], [18], [19]. However, multiplying the above bound by $(1-\alpha)$ on both sides and letting $\alpha \to 1$ yields ∞ on the right-hand side. Therefore, the above bound is not useful to understand the behavior of MDM in average-reward MDPs.

In the study of MDMs, two types of results have been obtained for discounted-reward MDPs: (i) global convergence in the tabular setting with perfect value function estimation, and (ii) performance bounds in the non-tabular setting with function approximations. In this paper, our main contribution is to show the analog of (i) for average-reward MDPs. For completeness, we also discuss the non-tabular setting by leveraging our recent results for approximate policy iteration in [20].

A. Related Work

Average reward MDPs have been well studied in the context of reinforcement learning [3], [21]–[26]. It is employed to model scenarios where the importance associated with the rewards does not decay with time. Unlike the discounted reward MDPs formulation, average reward MDPs do not possess a discount factor $\alpha < 1$ to aid in the convergence of standard learning and dynamic programming techniques [27], [28]. Hence in many situations, it is necessary to devise different proof techniques to study performance of well-known algorithms in the average reward setting [20], [29], [30].

Some of the very early works on average reward TD-Learning include [31], [32]. In [21], four natural actor-critic algorithms in the context of average reward are considered, and their asymptotic convergence to a neighbourhood around the local maxima is proven. In [33], actor critic methods with function approximation are studied, where the feature vectors are expressed as a span of functions of the parametrized class of policies. Episodic model free reinforcement learning algorithms are presented in [34]; however the episodic setting is very different from the setting we study in this paper. In [20], policy based reinforcement learning algorithms are modelled as variants of approximate policy iteration and the corresponding performance bounds are provided.

To the best of our knowledge, no global convergence results are available in the average reward case, even when the value function can be estimated precisely. On the other hand such global convergence results are available for NPG [35], [6] and for also for MDMs [11], [12] in the case of discounted reward MDPs .

The authors are with the Department of Electrical and Computer Engineering and the Coordinated Science Lab at the University of Illinois Urbana-Champaign, Urbana, IL 61801, USA.

The outline of the rest of the paper is as follows: Section II contains model and preliminaries where the relationship between average reward and discounted reward MDPs is also discussed. Section III contains the main body of the paper. It begins with the irreducibility assumption necessary for our proof of convergence of average reward MDMs, followed by the algorithm and its finite time convergence analysis. After a discussion on our assumption, Section III ends with the finite time bounds for non-tabular MDMs with function approximation in average reward MDPs. Section IV contains concluding remarks.

II. Model and Preliminaries

We consider infinite horizon Markov Decision Processes (MDPs) with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$ and class of randomized policies $\pi \in \Pi$. Let $\nabla(\mathcal{A})$ indicate the class of probability distributions over $\mathcal{A}$. Then $\Pi : \mathcal{S} \to \nabla(\mathcal{A})$. The underlying environment is modeled by a transition kernel $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \nabla(\mathcal{S})$. The probability of transition from s to s' under policy π is given by $\mathbb{P}(s'|s, \pi(s)) = \sum_{a\in\mathcal{A}} \pi(a|s)\mathbb{P}(s'|s, a)$. The single step reward associated with a policy π and state s is denoted by $r(s, \pi(s)) = \sum_{a\in\mathcal{A}} \pi(a|s)r(s, a)$. We assume the single step reward associated with any state and action is bounded.

The average reward associated with policy π is denoted by J^π and is defined as:

$$J^\pi = \lim_{T\to\infty} \frac{\mathbb{E}_\pi\left[\sum_{i=0}^{T-1} r(x_i, \pi(x_i))\right]}{T},$$

where x_i is the state at time i and the expectation is taken with respect to the transition kernel $\mathbb{P}_\pi$. Under standard assumptions [1], [2], J^π is independent of the distribution of the initial state x_0. Assuming π induces an ergodic Markov chain, J^π can be alternatively expressed as, $J^\pi = \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{a\in\mathcal{A}} \pi(a|s)r(s, a)$, where $d^\pi(s)$ is the stationary distribution over $\mathcal{S}$ under the policy π. The state-action value function $Q^\pi(s, a)$ associated with a policy π is defined as the solution to the average reward Bellman equation given by:

$$J^\pi \mathbb{1} + Q^\pi(s, a) = r(s, a) + \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s, a)Q^\pi(s', \pi(s')), \quad (1)$$

where $\mathbb{1}$ is the all ones vector. In order to obtain the state value function $V^\pi(s)$ associated with policy π, Equation (1) is averaged with the policy vector π to obtain,

$$J^\pi \mathbb{1} + V^\pi(s) = r(s, \pi(s)) + \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s, \pi(s))V^\pi(s'), \quad (2)$$

where $V^\pi(s') = \sum_{a\in\mathcal{A}} \pi(a|s')Q^\pi(s', a)$. We note that V^π is the relative value function, even though we refer to it as the value function for compactness. The advantage function A with respect to policies π and π', for all states $s \in \mathcal{S}$, is defined as

$$A^\pi(s, \pi'(s)) = Q^\pi(s, \pi'(s)) - V^\pi(s). \quad (3)$$

The advantage function $A^\pi(s, a)$ is analogous to the definition in the discounted-reward case [36], but the interpretation is a bit trickier in the average-reward case due to the fact that V^π is the relative value function. Let the optimal average reward be denoted as $J^{\pi^*} = \max_{\pi\in\Pi} J^\pi$ where $\pi^* = \operatorname{argmax}_{\pi\in\Pi} J^\pi$. Then under mild regularity conditions, there exists a value function V^{π^*} that satisfies the average reward Bellman optimality equation given by:

$$\begin{aligned} J^{\pi^*}\mathbb{1} + V^{\pi^*}(s) &= \max_{\pi\in\Pi} r(s, \pi(s)) + \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s, \pi(s))V^{\pi^*}(s') \\ &= r(s, \pi^*(s)) + \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s, \pi^*(s))V^{\pi^*}(s'). \end{aligned}$$

Note that the value functions Q^π and V^π corresponding to any policy π are unique up to an additive constant, by virtue of Equation (1) and Equation (2). In the next section we discuss the relationship between average reward MDPs and discounted reward MDPs.

A. Relationship to Discounted Reward MDP

The value function associated with a state $s \in \mathcal{S}$, policy π and discount factor $0 < \alpha < 1$ is defined as: $V^\alpha_\pi(s) = \mathbb{E}_\pi\left[\sum_{i=0}^\infty \alpha^i r(s_i, \pi(s_i))\middle| s_0 = s\right]$. Similarly, the state-action value function associated with (s, a) is defined as: $Q^\alpha_\pi(s, a) = \mathbb{E}_\pi\Big[r(s_0, a_0) + \sum_{i=1}^\infty \alpha^i r(s_i, \pi(s_i))\Big| s_0 = s, a_0 = a\Big]$. The advantage $A^\alpha_\pi(s, a)$ of using action a instead of policy π is defined as

$$A^\alpha_\pi(s, a) = Q^\alpha_\pi(s, a) - V^\alpha_\pi(s). \quad (4)$$

The associated discounted reward $J^\alpha_\pi(\eta)$ is the value function weighted with the initial distribution η over the state space, that is $J^\alpha_\pi(\eta) = \sum_{s\in\mathcal{S}} \eta(s)V^\alpha_\pi(s)$.

1) Mirror Descent-Based Policy methods: Since the objective of most reinforcement learning algorithms is to determine the optimal policy yielding maximum discounted reward, a popular algorithm to achieve that is natural policy gradient. NPG is closely related to mirror descent in the tabular setting whose objective at the kth iteration is to maximize:

$$\begin{aligned} \pi_{k+1} = \operatorname*{argmax}_{\pi\in\Pi} &\nabla^\intercal_\pi J^\alpha_{\pi_k}(\eta)(\pi - \pi_k) \\ &- \frac{1}{\beta}\sum_{s\in\mathcal{S}} d^{\pi_k}_\eta(s)\mathsf{D}_{\mathrm{KL}}\left(\pi(\cdot\|s)\|\pi_k(\cdot\|s)\right), \end{aligned} \quad (5)$$

where $d^{\pi_k}_\eta(s) = (1-\alpha)\sum_{i=0}^\infty \alpha^i \mathbb{P}_\pi(s_k = s|s_0 \sim \eta)$ is the normalized state visitation measure corresponding to state s and $\mathsf{D}_{\mathrm{KL}}\left(\pi(\cdot\|s)\|\pi_k(\cdot\|s)\right)$ is the Kullback-Leibler divergence between policies π and π_k. State visitation measure is the stationary distribution analog of the discounted reward MDP. Under tabular policy class, it can be shown that

$$\frac{\partial J^\alpha_\pi}{\partial\pi(a|s)} = \frac{1}{1-\alpha} d^\pi_\eta(s) A^\alpha_\pi(s, a). \quad (6)$$

From Equation (5) and Equation (6), we get the update equation

$$\pi_{k+1}(a|s) = \frac{\pi_k(a|s)e^{\beta A^\alpha_{\pi_k}(s,a)}}{\sum_{a'\in\mathcal{A}} \pi_k(a'|s)e^{\beta A^\alpha_{\pi_k}(s,a')}}. \quad (7)$$

Given the definition of the advantage function in Equation (4), the above update equation can be equivalently written as $\pi_{k+1}(a|s) = \frac{\pi_k(a|s)e^{\beta Q^{\alpha}_{\pi_k}(s,a)}}{\sum_{a' \in \mathcal{A}} \pi_k(a'|s) e^{\beta Q^{\alpha}_{\pi_k}(s,a')}}$. In the limit as $\beta \to \infty$, the policy obtained under such an update is identical to the one obtained through policy iteration. Hence MDM can be interpreted as soft policy iteration. A finite β ensures that all actions are explored with some non-zero probability which is crucial in order to effectively utilize algorithms such as TD learning to determine $Q^{\alpha}_{\pi}(s,a)$ and $A^{\alpha}_{\pi}(s,a)$.

Let $V^{\alpha}_{\pi*}(s) = \max_{\pi \in \Pi} V^{\alpha}_{\pi}(s)$. Assuming the initial policy π_0 is uniformly randomized across all actions and the bounded rewards (with a maximum value of 1), under the update rule Equation (7), previous literature has shown convergence of the policy updates to the optimal policy as below:

$$V^{\alpha}_{\pi^*}(s) \leq V^{\alpha}_{\pi_T}(s) + \frac{1}{T(1-\alpha)^3} + \frac{\log(|\mathcal{A}|)}{\beta T}, \qquad \forall s \in \mathcal{S}. \tag{8}$$

However this performance bound is vacuous in the context of average reward MDP as explained below.

2) Average Reward MDPs: Note that under mild regularity conditions standard MDP theory establishes the following relation between discounted reward MDPs and average reward MDPs:

$$J^{\pi} = \lim_{\alpha \to 1} (1-\alpha) V^{\alpha}_{\pi}(s). \tag{9}$$

Since the average cost is independent of the initial state, the above relation is true for all states $s \in \mathcal{S}$. Upon multiplying Equation (8) with $(1-\alpha)$ and setting $\alpha \to 1$, we still are left with a $(1-\alpha)^2$ in the denominator on the RHS, which leaves us with vacuous bounds for convergence of MDM for average reward MDPs. But since this is only an upper bound, it is interesting to understand whether the algorithm converges in the average-reward case. In the next section, we outline under what conditions we obtain convergence of MDM for average reward MDPs as well, and subsequently provide a proof of the same.

III. Average Reward MDM

We now present the average reward MDM in the tabular setting with perfect estimates of the value functions. The algorithm is a natural extension of the corresponding algorithm in the discounted-reward case.

Algorithm 1 Average Reward MDM

Input: $\beta > 0,\ \pi_0(\cdot|s) \in \Delta(\mathcal{A}),\ \forall s \in \mathcal{S}$

for $k = 0, \ldots, T-1$

1: Compute $A^{\pi_k}(s,a)\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$

2: Update for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

$$\pi_{k+1}(a|s) = \frac{\pi_k(a|s) e^{\beta A^{\pi_k}(s,a)}}{\sum_{a' \in \mathcal{A}} \pi_k(a'|s) e^{\beta A^{\pi_k}(s,a')}} \tag{10}$$

endfor

Output: π_T

We prove the global convergence of the above algorithm under the following assumption.

Assumption 1: Let $d^{\pi}(s)$ be the stationary probability of being in state s under $\mathbb{P}_{\pi}$. Let π^* be the optimal policy. We assume that

$$\Delta = \inf_{\substack{\pi \in \Pi \\ s \in \mathcal{S}}} \frac{d^{\pi}(s)}{d^{\pi^*}(s)} > 0. \tag{11}$$

$\diamond$

Later, we will discuss how most practical MDPs can be made to satisfy this assumption for a small loss in performance.

A. Convergence Analysis of Algorithm 1

Our convergence analysis follows the same outline as in the case of discounted-reward MDPs: we first utilize the average reward performance-difference lemma, then show the monotonicity of the average reward and finally, we use a weighted KL distance as a Lyapunov function to establish convergence. However, each of the steps has to be appropriately modified to get rid of the dependence on the discount factor and replace it with other relevant quantities.

Lemma 1. *(Performance Difference Lemma) Recall the definition of advantage function in Equation* (3)*. Let J^{π} and $J^{\pi'}$ be the average rewards associated with policies π and π' respectively. Then it is true that,*

$$J^{\pi} - J^{\pi'} = \sum_{s \in \mathcal{S}} d^{\pi}(s) A^{\pi'}(s, \pi(s)), \tag{12}$$

where d^{π} is the stationary distribution induced over $\mathcal{S}$ by policy π.

Proof. The proof can be found in [37]. $\square$

As in the discounted-reward case, the performance difference lemma plays a key role in the proof of convergence of average reward MDM, especially to establish the monotonicity of the sequence of average rewards obtained through Algorithm 1. However, in our average-reward case, it is also utilized to study the drift of the Lyapunov function used to prove the convergence of the algorithm.

1) Monotonicity of the MDM Update: We now present a key lemma used to prove the convergence of the MDM update to the optimal policy.

Lemma 2. *Given π_k which are generated according to Equation* (10)*, the corresponding sequence of average rewards J^{π_k} are increasing, that is,*

$$J^{\pi_{k+1}} - J^{\pi_k} \geq 0. \tag{13}$$

Proof. From the MDM update Equation (10) we have the following,

$$A^{\pi_k}(s,a) = \frac{1}{\beta} \log \left(\frac{z_k(s) \pi_{k+1}(a|s)}{\pi_k(a|s)} \right).$$

From Lemma 1, we know that

$$\begin{aligned}
J^{\pi_{k+1}}-J^{\pi_k} &= \sum_{s\in\mathcal{S}} d^{\pi_{k+1}}(s)\sum_{a\in\mathcal{A}}\pi_{k+1}(a|s)A^{\pi_k}(s,a)\\
&= \frac{1}{\beta}\sum_{s\in\mathcal{S}} d^{\pi_{k+1}}(s)\sum_{a\in\mathcal{A}}\pi_{k+1}(a|s)\log\left(\frac{z_k\pi_{k+1}(a|s)}{\pi_k(a|s)}\right)\\
&= \frac{1}{\beta}\sum_{s\in\mathcal{S}} d^{\pi_{k+1}}(s)\underbrace{\sum_{a\in\mathcal{A}}\pi_{k+1}(a|s)\log\left(\frac{\pi_{k+1}(a|s)}{\pi_k(a|s)}\right)}_{\mathsf{D}_{\mathrm{KL}}\left(\pi_{k+1}(\cdot|s)\|\pi_k(\cdot|s)\right)\geq 0}\\
&\quad+\frac{1}{\beta}\sum_{s\in\mathcal{S}} d^{\pi_{k+1}}(s)\sum_{a\in\mathcal{A}}\pi_{k+1}(a|s)\log\left(z_k(s)\right)
\end{aligned}$$

$$\geq \frac{1}{\beta}\sum_{s\in\mathcal{S}} d^{\pi_{k+1}}(s)\sum_{a\in\mathcal{A}}\pi_{k+1}(a|s)\log\left(\sum_{a'\in\mathcal{A}}\pi_k(a'|s)e^{\beta A^{\pi_k}(s,a')}\right)$$

$$\begin{aligned}
&\overset{(a)}{\geq} \frac{1}{\beta}\sum_{s\in\mathcal{S}} d^{\pi_{k+1}}(s)\underbrace{\sum_{a\in\mathcal{A}}\pi_{k+1}(a|s)}_{=1}\sum_{a'\in\mathcal{A}}\pi_k(a'|s)\beta A^{\pi_k}(s,a')\\
&= \sum_{s\in\mathcal{S}} d^{\pi_{k+1}}(s)\sum_{a'\in\mathcal{A}}\pi_k(a'|s)A^{\pi_k}(s,a').
\end{aligned}$$

where (a) is due to Jensen's inequality and concavity of the log function. Substituting for $A^{\pi_k}(s,a')$,

$$J^{\pi_{k+1}}-J^{\pi_k}\geq\sum_{s\in\mathcal{S}} d^{\pi_{k+1}}(s)\sum_{a'\in\mathcal{A}}\pi_k(a'|s)(Q^{\pi_k}(s,a')-V^{\pi_k}(s))$$

Since $V^{\pi_k}(s)=\sum_{a'\in\mathcal{A}}\pi_k(a'|s)Q^{\pi_k}(s,a')$, we obtain

$$J^{\pi_{k+1}}-J^{\pi_k}\geq 0.$$

□

Now that we have established the monotonicity of the average reward associated with the policy iterates, we now compare the policy iterates with the optimal policy. In order to prove the convergence of the MDM iterates, consider the following Lyapunov function whose argument is a policy vector.

$$W(\pi)=\sum_{s\in\mathcal{S}} d^{\pi^*}(s)\mathsf{D}_{\mathrm{KL}}\left(\pi^*(\cdot|s)\|\pi(\cdot|s)\right). \tag{14}$$

The idea is to show that with the subsequent iterates, the value of the associated Lyapunov function reduces, which implies that the NPG iterates approach the optimal policy. A lemma important to prove the result is stated below.

Lemma 3. *Consider the sequence of policies π_k obtained from Algorithm 1. It is true that*

$$W(\pi_{k+1})-W(\pi_k)=\sum_{s\in\mathcal{S}} d^{\pi^*}(s)\log(z_k(s))-\beta\left(J^{\pi^*}-J^{\pi_k}\right),$$

where $z_k(s)=\sum_{a'\in\mathcal{A}}\pi_k(a'|s)e^{\beta A^{\pi_k}(s,a')}$.

Proof. From Equation (14) we know that,

$$W(\pi_{k+1})-W(\pi_k)=\sum_{s\in\mathcal{S}} d^{\pi^*}(s)\sum_{a\in\mathcal{A}}\pi^*(a|s)\log\left(\frac{\pi_k(a|s)}{\pi_{k+1}(a|s)}\right).$$

Using the update Equation (10),

$$\begin{aligned}
W(\pi_{k+1})-W(\pi_k) &= \sum_{s\in\mathcal{S}} d^{\pi^*}(s)\sum_{a\in\mathcal{A}}\pi^*(a|s)\log\left(\frac{z_k(s)}{e^{\beta A^{\pi_k}(s,a)}}\right)\\
&= \sum_{s\in\mathcal{S}} d^{\pi^*}(s)\sum_{a\in\mathcal{A}}\pi^*(a|s)\log\left(z_k(s)\right)\\
&\quad-\beta\sum_{s\in\mathcal{S}} d^{\pi^*}(s)\sum_{a\in\mathcal{A}}\pi^*(a|s)A^{\pi_k}(s,a)\\
&= \sum_{s\in\mathcal{S}} d^{\pi^*}(s)\log\left(z_k(s)\right)-\beta\left(J^{\pi^*}-J^{\pi_k}\right),
\end{aligned}$$

where the last step follows from Lemma 1. □

The above lemma captures the impact of consecutive policy iterates on the chosen Lyapunov function. The main result follows by telescoping the expression from the previous lemma as proved in the theorem below.

Theorem 4. *The sequence of policies π_k generated through the update rule in Equation* (10)*, correspond to MDPs with average reward J^{π_k}, which approach the optimal average reward as below:*

$$J^{\pi^*}-J^{\pi_T}\leq\frac{W(\pi_0)-W(\pi_T)}{\beta T}+\frac{J^{\pi_T}-J^{\pi_0}}{\Delta T}.$$

Proof. From Lemma 1,

$$\begin{aligned}
J^{\pi_{k+1}}-J^{\pi_k} &= \sum_{s\in\mathcal{S}} d^{\pi_{k+1}}(s)A^{\pi_k}(s,\pi_{k+1}(s))\\
&= \frac{1}{\beta}\Bigg(\sum_{s\in\mathcal{S}} d^{\pi_{k+1}}(s)\underbrace{\sum_{a\in\mathcal{A}}\pi_{k+1}(a|s)\log\left(\frac{\pi_{k+1}(a|s)}{\pi_k(a|s}\right)}_{\mathsf{D}_{\mathrm{KL}}\left(\pi_{k+1}(\cdot|s)\|\pi_k(\cdot|s)\right)\geq 0}\\
&\quad+\sum_{s\in\mathcal{S}} d^{\pi_{k+1}}(s)\underbrace{\sum_{a\in\mathcal{A}}\pi_{k+1}(a|s)}_{=1}\log(z_k(s))\Bigg)\\
&\geq \frac{1}{\beta}\sum_{s\in\mathcal{S}} d^{\pi^*}(s)\frac{d^{\pi_{k+1}}(s)}{d^{\pi^*}(s)}\log\left(z_k(s)\right)\\
&\overset{(a)}{\geq} \frac{\Delta}{\beta}\sum_{s\in\mathcal{S}} d^{\pi^*}(s)\log\left(z_k(s)\right).
\end{aligned}$$

where (a) follows from Section III. From Lemma 3,

$$J^{\pi_{k+1}}-J^{\pi_k}\geq\frac{\Delta}{\beta}\left(W(\pi_{k+1})-W(\pi_k)+\beta\left(J^{\pi^*}-J^{\pi_k}\right)\right).$$

Iterating the above expression for $k=0,\ldots,T$ yields,

$$J^{\pi_T}-J^{\pi_0}\geq\frac{\Delta}{\beta}\left(W(\pi_T)-W(\pi_0)+\beta\left(TJ^{\pi^*}-\sum_{i=0}^{T-1}J^{\pi_i}\right)\right).$$

$$J^{\pi^*}-\frac{\sum_{i=0}^{T-1}J^{\pi_i}}{T}\leq\frac{W(\pi_0)-W(\pi_T)}{\beta T}+\frac{J^{\pi_T}-J^{\pi_0}}{\Delta T}.$$

From Lemma 2, we know that the sequence of policies obtained correspond to MDPs with monotonically increasing average rewards, that is $J^{\pi_{k+1}}\geq J^{\pi_k}$. Hence, we obtain,

$$J^{\pi^*}-J^{\pi_T}\leq\frac{W(\pi_0)-W(\pi_T)}{\beta T}+\frac{J^{\pi_T}-J^{\pi_0}}{\Delta T}.$$

□

B. Discussion: Assumption 1

Assumption 1 requires the stationary distribution induced over the state space by all policies within the randomized policy function class to be uniformly bounded away from zero. Depending on the underlying transition kernel, this assumption need not always be satisfied. However, it is often the case that there exists a policy (possibly randomized) under which the resulting Markov chain is irreducible. An example of such a policy is one where each action is chosen uniformly at random in each state. Let us denote such a policy by π'. We can modify the original MDP as follows: in each state, with probability ϵ, pick an action specified by policy π' and with probability $1-\epsilon$, choose an action specified by another policy π. Now, the MDP can be viewed as optimizing over policy π with π' being fixed. This induces an $O(\epsilon)$ suboptimality in the average reward that MDM converges to and thus, our convergence result should be interpreted as a near optimality result.

C. Discussion: Non-Tabular Average Reward MDM

While performance bounds for non-tabular average-reward MDM have not been provided in prior literature, they can be easily inferred from the results in [20], [38].

Algorithm 2 Non Tabular Average Reward MDM

Input: $\beta > 0,\ \pi_0(\cdot|s) \in \Delta(\mathcal{A}),\ \forall s \in \mathcal{S}$

for $k = 0, \ldots, T-1$

1: Use TD-Learning to compute $Q_k(s,a)$ as an approximation to $Q^{\pi_k}(s,a)\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$

2: Update for all $(s,a) \in \mathcal{S} \times \mathcal{A}$

$$\pi_{k+1}(a|s) = \frac{\pi_k(a|s) e^{\beta Q_k(s,a)}}{\sum_{a' \in \mathcal{A}} \pi_k(a'|s) e^{\beta Q_k(s,a')}} \tag{15}$$

endfor

Output: π_T

Note that in general TD-Learning algorithms are used to learn the state-action value function $Q^{\pi_k}(s,a)$. However, since we know the policy π_k, one can average the learnt approximation Q_k to obtain V_k and consequently the advantage function A_k as given by Equation (3).

TD Learning algorithms in the context of average reward MDPs are presented in [30], where function approximation is utilized for state-action value function $Q^{\pi_k}(s,a)$. More precisely, $Q_k(s,a) = \phi(s,a)^\intercal \theta_k$ where $\phi(s,a) \in \mathbb{R}^d$ is the feature vector, and θ_k is the parameter vector that requires estimation at every iteration. The TD learning error at each time step can then be expressed as follows:

$$\|Q_k - Q^{\pi_k}\|_\infty \leq \underbrace{\|\Phi(\theta_k - \theta_k^*)\|_\infty}_{\textit{policy evaluation error}} + \underbrace{\|\Phi\theta_k^* - Q^{\pi_k}\|_\infty}_{\textit{function approximation error}}, \tag{16}$$

where $\theta_k^* = \text{argmin}_{\theta \in \mathbb{R}^d} \|\Phi\theta - Q^{\pi_k}\|_2^2$. The policy evaluation error depends on the number of samples utilized to learn the best feature vector θ_k whereas the function approximation error is a function of the richness of the feature vector Φ (and does not depend on the number of samples).

Given a policy π_k and state s, since $V_k(s)$ does not depend on action, the policy update equation in Algorithm 2 is equivalent to Equation (10).

Define the Bellman operator $\mathsf{T}_\pi : \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ with respect to a policy π and optimal Bellman operator $\mathsf{T} : \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ as follows: $(\mathsf{T}_\pi Q)(s,a) = r(s,a) + \sum_{s' \in \mathcal{S}} \widehat{P}_\pi(s',a'|s,a) Q(s',a')$ and $(\mathsf{T}Q)(s,a) = r(s,a) + \max_{a' \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \widehat{P}(s',a'|s,a) Q(s',a')$ The policy improvement error obtained as a result of using MDM policy update π_k as opposed to the action maximizing $\mathsf{T}Q_k$ is given by:

$$0 \leq (\mathsf{T}Q_k - \mathsf{T}_{\pi_{k+1}} Q_k)(s,a) \leq \frac{1}{\beta} \log \frac{1}{\min_{s \in \mathcal{S}} \pi_{k+1}(a^*(s)|s)} \tag{17}$$

where $a^*(s)$ is the optimal action at state s obtained as the maximizing argument of the optimal Bellman operator acting on Q_k. The error is quantified in terms of the probability of choosing the optimal action under the policy obtained through MDM. The details of the above bound can be found in [38]. Let J^{π^*} be the optimal average reward associated with $\widehat{\mathbb{P}}$. Given the policy evaluation error in Equation (16) and the policy improvement error in Equation (17) as a result of the MDM update π_k given by Equation (15), we obtain the following finite time performance bound:

$$\begin{aligned}
\mathbb{E}\big[J^{\pi^*} - J_{\pi_{T+1}}\big] \leq & \underbrace{(1-\gamma)^T\, \mathbb{E}\Big[J^{\pi^*} - \min_i \left(\mathsf{T}Q_0 - Q_0\right)(i)\Big]}_{\textit{Initial condition error}} \\
& + \underbrace{\sum_{\ell=1}^{T-1} (1-\gamma)^{\ell-1}\, \mathbb{E}\big[\big\|\mathsf{T}_{\pi_{T+1-\ell}} Q_{T-\ell} - \mathsf{T}Q_{T-\ell}\big\|_\infty\big]}_{\textit{Policy improvement error}} \\
& + \underbrace{\mathbb{E}\left[\big\|\mathsf{T}_{\pi_{T+1}} Q_T - \mathsf{T}Q_T\big\|_\infty\right]}_{\textit{Policy improvement error}} \\
& + \underbrace{2 \sum_{\ell=0}^{T-1} (1-\gamma)^\ell\, \mathbb{E}\big[\big\|Q_{T-\ell} - Q_{\pi_{T-\ell}}\big\|_\infty\big]}_{\textit{Policy evaluation error}}.
\end{aligned}$$

where $\gamma = \inf_{\substack{s \in \mathcal{S} \\ \pi \in \Pi}} d^\pi(s)$ where $d^\pi(s)$ is the stationary measure associated with s under policy π. More details on this bound can be found in [20].

Note that in the tabular case, with perfect value function information, we were able to prove global convergence. However, in the non tabular version of average reward MDM, relative value function estimates are obtained through TD-Learning, an algorithm that requires each (state, action) pair to be visited infinitely often. As $\beta \to \infty$, it is easy to see that some state action pairs will never be visited, which leads to a significant error in their relative value function estimation. However, from Equation (17), we know that as $\beta \to \infty$, the policy improvement error associated with the MDM update tends to zero. Hence the performance bound for non-tabular average reward MDM crucially depends on the choice of β, i.e., there is a tradeoff between policy improvement error and policy evaluation error. This raises an interesting question about the learning algorithms utilized for the purpose of relative value function evaluation. If there is a way to estimate the relative function in a manner that is independent of β, then it may be possible to get

better performance bounds depending upon the choice of the behavioral policy. This is an interesting question to explore in the future.

IV. Conclusion

In this paper we prove the global convergence of MDM, along with finite time bounds, in the tabular setting of average reward MDPs with perfect value function estimates. We then proceed to leverage some results from recent literature to present finite time performance bounds for non-tabular MDM with approximate policy evaluations and policy improvements. This work extends some of the very well known core results in the context of discounted reward MDPs to the average reward MDP domain, thus opening up more avenues for research in average reward MDMs.

We note that one can use connection between learning from expert advice and tabular MDM as in [39] to show that the difference between the average reward and the reward obtained by MDM decays as $O(1/\sqrt{T})$, instead of $O(1/T)$ as in our paper. We note that both results are useful in different respects: while we establish a faster rate of convergence in terms of T, the result in [39] is independent of Δ and thus, independent of the size of the state space.

References

[1] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[2] D. Bertsekas, *Dynamic programming and optimal control: Volume I*, vol. 1. Athena scientific, 2012.

[3] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[4] S. M. Kakade, "A natural policy gradient," *Advances in neural information processing systems*, vol. 14, 2001.

[5] Y. Liu, K. Zhang, T. Basar, and W. Yin, "An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7624–7636, 2020.

[6] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 4431–4506, 2021.

[7] A. S. Nemirovskij and D. B. Yudin, "Problem complexity and method efficiency in optimization," 1983.

[8] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, "Fast global convergence of natural policy gradient methods with entropy regularization," *Operations Research*, vol. 70, no. 4, pp. 2563–2578, 2022.

[9] W. Zhan, S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi, "Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence," *arXiv preprint arXiv:2105.11066*, 2021.

[10] S. Cen, F. Chen, and Y. Chi, "Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 2833–2838, IEEE, 2022.

[11] Y. Li and G. Lan, "Policy mirror descent inherently explores action space," *arXiv preprint arXiv:2303.04386*, 2023.

[12] G. Lan, "Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes," *Mathematical programming*, vol. 198, no. 1, pp. 1059–1106, 2023.

[13] Y. Li, T. Zhao, and G. Lan, "Robust policy mirror descent for controlling uncertain markov decision process," 2022.

[14] Y. Li, T. Zhao, and G. Lan, "Homotopic policy mirror descent: Policy convergence, implicit regularization, and improved sample complexity," *arXiv preprint arXiv:2201.09457*, 2022.

[15] Y. Jin and A. Sidford, "Efficiently solving mdps with stochastic mirror descent," in *International Conference on Machine Learning*, pp. 4890–4900, PMLR, 2020.

[16] W. H. Montgomery and S. Levine, "Guided policy search via approximate mirror descent," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[17] M. Geist, B. Scherrer, and O. Pietquin, "A theory of regularized markov decision processes," in *International Conference on Machine Learning*, pp. 2160–2169, PMLR, 2019.

[18] S. M. Ross, *Applied probability models with optimization applications*. Courier Corporation, 2013.

[19] D. Blackwell, "Discrete dynamic programming," *The Annals of Mathematical Statistics*, pp. 719–726, 1962.

[20] Y. Murthy, M. Moharrami, and R. Srikant, "Performance bounds for policy-based average reward reinforcement learning algorithms," *arXiv preprint arXiv:2302.01450*, 2023.

[21] S. Bhatnagar, R. S. Sutton, M. Ghavamzadeh, and M. Lee, "Natural actor–critic algorithms," *Automatica*, vol. 45, no. 11, pp. 2471–2482, 2009.

[22] P. Thomas, "Bias in natural actor-critic algorithms," in *International conference on machine learning*, pp. 441–448, PMLR, 2014.

[23] S. Mahadevan, "Average reward reinforcement learning: Foundations, algorithms, and empirical results," *Recent advances in reinforcement Learning*, pp. 159–195, 1996.

[24] M. Ghavamzadeh and S. Mahadevan, "Hierarchical average reward reinforcement learning.," *Journal of Machine Learning Research*, vol. 8, no. 11, 2007.

[25] A. Schwartz, "A reinforcement learning method for maximizing undiscounted rewards," in *Proceedings of the tenth international conference on machine learning*, vol. 298, pp. 298–305, 1993.

[26] S. P. Singh, "Reinforcement learning algorithms for average-payoff markovian decision processes," in *AAAI*, vol. 94, pp. 700–705, 1994.

[27] R. A. Howard, "Dynamic programming and markov processes.," 1960.

[28] D. J. White, "Dynamic programming, markov chains, and the method of successive approximations," *Journal of Mathematical Analysis and Applications*, vol. 6, no. 3, pp. 373–376, 1963.

[29] J. Van der Wal, "The method of value oriented successive approximations for the average reward markov decision process," *Operations-Research-Spektrum*, vol. 1, no. 4, pp. 233–242, 1980.

[30] S. Zhang, Z. Zhang, and S. T. Maguluri, "Finite sample analysis of average-reward td learning and q-learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1230–1242, 2021.

[31] J. N. Tsitsiklis and B. Van Roy, "Average cost temporal-difference learning," *Automatica*, vol. 35, no. 11, pp. 1799–1808, 1999.

[32] J. N. Tsitsiklis and B. Van Roy, "On average versus discounted reward temporal-difference learning," *Machine Learning*, vol. 49, no. 2-3, pp. 179–191, 2002.

[33] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," *Advances in neural information processing systems*, vol. 12, 1999.

[34] C.-Y. Wei, M. J. Jahromi, H. Luo, H. Sharma, and R. Jain, "Model-free reinforcement learning in infinite-horizon average-reward markov decision processes," in *International conference on machine learning*, pp. 10170–10180, PMLR, 2020.

[35] S. Khodadadian, P. R. Jhunjhunwala, S. M. Varma, and S. T. Maguluri, "On the linear convergence of natural policy gradient algorithm," in *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 3794–3799, IEEE, 2021.

[36] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, "Reinforcement learning: Theory and algorithms," *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pp. 10–4, 2019.

[37] X.-R. Cao, "Single sample path-based optimization of markov chains," *Journal of optimization theory and applications*, vol. 100, pp. 527–548, 1999.

[38] Z. Chen and S. T. Maguluri, "Sample complexity of policy-based methods under off-policy sampling and linear function approximation," in *International Conference on Artificial Intelligence and Statistics*, pp. 11195–11214, PMLR, 2022.

[39] E. Even-Dar, S. M. Kakade, and Y. Mansour, "Online markov decision processes," *Mathematics of Operations Research*, vol. 34, no. 3, pp. 726–736, 2009.