Analysis of Stochastic Processes through Replay Buffers

Shirli Di Castro Shashua [1] **Shie Mannor** [2] **Dotan Di Castro** [3]

Abstract

Replay buffers are a key component in many reinforcement learning schemes. Yet, their theoretical properties are not fully understood. In this paper we analyze a system where a stochastic process X is pushed into a replay buffer and then randomly sampled to generate a stochastic process Y from the replay buffer. We provide an analysis of the properties of the sampled process such as stationarity, Markovity and autocorrelation in terms of the properties of the original process. Our theoretical analysis sheds light on why replay buffer may be a good de-correlator. Our analysis provides theoretical tools for proving the convergence of replay buffer based algorithms which are prevalent in reinforcement learning schemes.

1. Introduction

A Replay buffer (RB) is a mechanism for saving past generated data samples and for sampling data for off-policy reinforcement learning (RL) algorithms (Lin, 1993). The RB serves a First-In-First-Out (FIFO) buffer with a fixed capacity and it enables sampling mini-batches from previously saved data points. Its structure and sampling mechanism provide a unique characteristic: the RB serves as *de-correlator* of data samples. Typically, the agent in RL algorithms encounters sequences of highly correlated states and learning from these correlated data points may be problematic since many deep learning algorithms suffer from high estimation variance when data samples are dependent. Thus, a mechanism that decorrelates the input such as the RB can improve data efficiency and reduce sample complexity.

Since its usage in the DQN algorithm (Mnih et al., 2013), RB mechanism have become popular in many off-policy RL algorithms (Lillicrap et al., 2015; Haarnoja et al., 2018). Previous work has been done on the empirical benefits of RB usage (Fedus et al., 2020; Zhang & Sutton, 2017), but still there is a lack in theoretical understanding of how the RB mechanism works. Understanding the properties of RBs is crucial for convergence and finite sample analysis of algorithms that use a RB in training. For the best of our knowledge, this is the first work to tackle these theoretical aspects.

In this work we focus on the following setup. We define a random process X that is pushed into a N samples size RB and analyze the characteristics of the stochastic process of K samples that is sampled from the RB. We analyze if properties of the original random process such as Markovity and stationarity are maintained and quantify the auto-correlation and covariance in the new RB process (later denoted by Y) when possible.

Our motivation comes from RL algorithms that use RB. Specifically, we focus on the induced Markov chain given a policy but we note that the analysis in this paper is also relevant to general random processes that are kept in a FIFO queue. This is relevant for domains such as First Come First Served domains (Laguna & Marklund, 2013). Our goal is to provide analytical tools for analyzing algorithms that use RBs. Our results can provide theoretical understanding of phenomena seen in experiments using RBs that have never been analyzed theoretically before. Our theory for RBs provides tools for proving convergence of RB-based RL algorithms.

Our main contributions are:

1. Formulating RBs as random processes and analyze their properties such as stationarity, Markovity, ergodicity, auto-correlation and covariance.
2. Comparing between properties of the original random process that was pushed into the RB and the sampled process at the output of the RB. Particularly we prove that when sampling uniformly from the RB, the RB forms as a de-correlator between the sampled batches.
3. Connecting our RB theory to RL by demonstrating this connection through a RB-based actor critic RL algorithm that samples K transitions from RB with size N for updating its parameters. We prove, for the first time, the asymptotic convergence of such RB-based actor critic algorithm.

[*]Equal contribution [1]Technion Institute of Technology, Haifa, Israel [2]Technion and NVIDIA Research, Israel [3]Bosch Center of AI, Haifa, Israel. Correspondence to: Shirli Di Castro Shashua <sdicastro@gmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

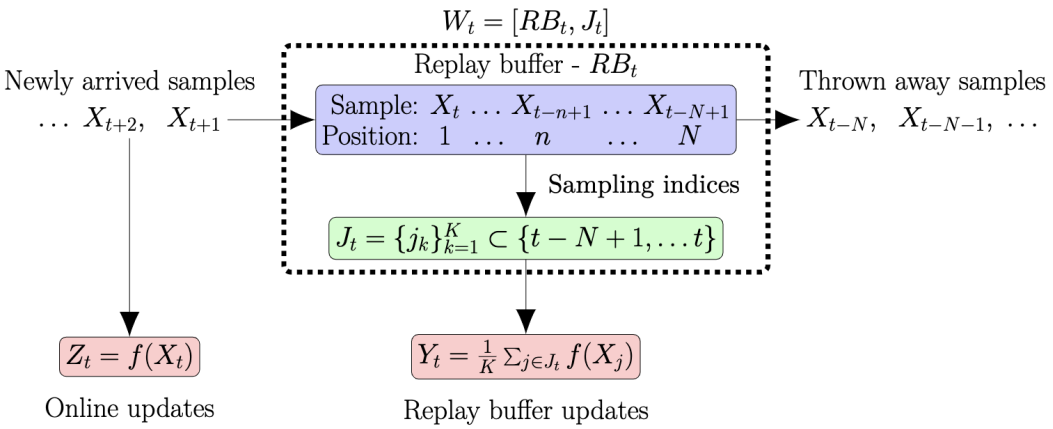

Figure 1. Replay buffer flow diagram: Process X enters the RB which stores $\{X_t, \ldots X_{t-N-1}\}$ in positions $(1, \ldots, N)$, respectively. As time proceeds and $t > N$, old transition are thrown away from the RB. At each time step t, a random subset of K time steps is sampled from the RB and is denoted as J_t. W is simply $[RB, J]$. Y is the process of averaging a function over X at times from the subset J. Lastly, the process Z is simply a function applied on the variable X. Comparing Y to Z, we can see that Z can serve as an online update while Y can serve as a RB-based update.

The paper is structured as follows. We begin with presenting the setup in Section 2. We then state our main results regarding RB properties in Section 3. In Section 4 we connect between our RB theory and its use in RL and provide a convergence proof for an RB-based actor critic algorithm. Afterward, in Section 5 we position our work in existing literature and conclude in Section 6.

2. Setup for Replay Buffer Analysis

2.1. Replay Buffer Structure

Let $X \triangleq (X_t)_{t=0}^{\infty}$ be a stochastic process where the subscript t indicates time. The samples are dynamically pushed into a Replay Buffer (RB; Lin, 1993; Mnih et al., 2013) of capacity N, i.e., it is a First-In-First-Out (FIFO) buffer that can hold the N latest samples. We define the state of the RB at time t with $RB_t = \{X_{t-N+1}, \ldots, X_t\}$. Suppose that the buffer cells are numbered from 1 to N. The latest observation of X is pushed into cell 1, the observation before into cell 2, etc. When a new observation arrives, the observation in cell n is pushed into cell $n+1$ for $1 \leq n < N$, while the observation in cell N is thrown away.

The random process $RB = (RB_t)_{t=0}^{\infty}$ contains the last N samples of X. The random process Y is defined as the average of random K samples (without replacement) out of the N samples and applying a function $f(\cdot) : \cdot \to \mathbb{R}^D$ [1] where D is the dimension of the algorithm[2]. The function $f(\cdot)$ may correspond to a typical RL function that one usually find in RL algorithms such as linear function approximation, Temporal Difference, etc. (Bertsekas & Tsitsiklis, 1996). We elaborate on possible RL functions in Section 4.2.

2.2. Replay Buffer Sampling Method

We analyze the "unordered sampling without replacement" strategy from the RB. We note that other sampling methods may be analyzed, but we chose this specific sampling due to its popularity in many deep reinforcement learning algorithms[3]. Let $\mathbb{N}$ be a set of indices: $\mathbb{N} = \{1, \ldots, N\}$ and let $\bar{J}$ be a subset of K indices from $\mathbb{N}$. Given N and K, let $\mathbb{C}_{N,K}$ be the set of all possible subsets $\bar{J}$ for specific N and K. Then, the probability of sampling subset $\bar{J}$ is $p_{\text{binom}}^{N,K}(\bar{J}) = \frac{1}{\binom{N}{K}}\ \forall \bar{J} \in \mathbb{C}_{N,K}$, where $\binom{N}{K} \triangleq \frac{N!}{(N-K)!K!}$ is the binomial coefficient.

2.3. Replay Buffer Related Processes

We denote the set of K temporal indices of the samples from RB by the random process J (corresponds to a *"Batch"* in Deep Learning) where $J_t = \{j_k\}_{k=1}^K \subset \{t-N+1, \ldots, t\}$[4]. Similarly, the corresponding K RB indices process is $\bar{J}$ where $\bar{J}_t \subset \{1, \ldots, N\}$. We remark that both J_t and $\bar{J}_t$ contain identical information but one refers to the absolute time, and one to the indices of the RB. We define the random process $W_t \triangleq [RB_t, J_t]$ which holds both the information

[1] We note that $f(\cdot)$ may also depend on t but we leave that for the sake of simplicity.

[2] For example, in linear function approximation of Actor-Critic algorithms, D is the dimension of the linear basis used to approximate the value function by the critic.

[3] In Section 6 we discuss shortly future directions for other sampling schemes.

[4] We note that in the first K steps the batch is of size smaller than K and in the first N steps the RB is not full.

on the RB as well on the sampling from it. For later usage, we define the process X_t going through a function $f(\cdot)$ with $Z_t \triangleq f(X_t)$. The resulting Y_t has the structure of

$$Y_t = \frac{1}{K}\sum_{j\in J_t} Z_j = \frac{1}{K}\sum_{j\in J_t} f(X_j).$$

The stochastic processes relations that are described above are visualized in Figure 1.

3. Replay Buffer Properties

In this section we analyze the properties of a random process Y that is sampled from the RB and used in some RL algorithm. Specifically, we analyze stationarity, Markovity, ergodicity, auto-correlation, and covariance.

3.1. Stationarity, Markovity and Ergodicity

The following Lemmas characterize the connection between different properties of X that enter the RB and the properties of the processes RB and Y.

Stationarity is not a typical desired RL property since we constantly thrive to improve the policy (and thus the induced policy) but we bring it here for the sake of completeness.

Lemma 1 (Stationarity)**.** *Let X_t and J_t be stationary processes. Then, RB_t and Y_t are stationary.*

The proof for Lemma 1 is deferred to Section A.1 in the supplementary material.

In the next Lemma we analyze when the process RB is Markovian. This property is important in RL analysis. However, we note that Y_t is not necessarily Markovian but W_t is Markovian. For this, we define $X_{n_1}^{n_2}(t) \triangleq \{X_{t-n_2+1},\ldots,X_{t-n_1+1}|RB_t\}$ as the set of random variables from process X stored in the RB at time t from position n_1 to n_2.

Lemma 2 (Markovity)**.** *Let X_t be a Markov process. Then: (1) RB_t and W_t are Markovian. (2) The transition probabilities of RB_t for $t \geq N$ are:*

$$P\left(RB_{t+1}|RB_t\right) = \begin{cases} P\left(X_{t+1}|X_t\right) & \text{if } X_t \in X_1^1(t) \\ & \text{and } RB_{t+1} = \{X_{t+1}\}\cup X_1^{N-1}(t), \\ 0 & \text{otherwise.} \end{cases}$$

If J_t is sampled according to "unordered sampling without replacement", then the transition probabilities of W_t for $t \geq N$ are:

$$P\left(W_{t+1}|W_t\right) = \begin{cases} \frac{1}{\binom{N}{K}} P\left(X_{t+1}|X_t\right) & \forall J_{t+1}\in\mathbb{C}_{N,K},\ \text{if } X_t \in X_1^1(t) \\ & \text{and } RB_{t+1} = \{X_{t+1}\}\cup X_1^{N-1}(t) \\ 0 & \text{otherwise.} \end{cases}$$

The proof for Lemma 2 is deferred to Section A.2 in the supplementary material.

In RL, the properties of apreiodicity and irreducible (that together form ergodicity; Norris, 1998) are crucial in many convergence proofs. The following states that these properties are preserved when using RB.

Lemma 3 (Ergodicity)**.** *Let X_t be a Markov process that is aperiodic and irreducible. Then, RB_t and W_t are aperiodic and irreducible. Moreover, every point $y \in supp(Y_t)$ is visited infinitely often.*

The proof for Lemma 3 is deferred to Section A.3 in the supplementary material.

3.2. Auto-Correlation and Covariance

In this section we analyze the auto-correlation and covariance of the process Y expressed by process X properties. When X is stationary, the auto-correlation and covariance functions for X are:

$$R_X(\tau) = \mathbb{E}[X_t X_{t+\tau}]$$
$$C_X(\tau) = \mathbb{E}\left[\left(X_t - \mathbb{E}\left[X_t\right]\right)\left(X_{t+\tau} - \mathbb{E}\left[X_{t+\tau}\right]\right)\right]$$

In the same way, the definition of the auto-correlation and covariance functions for the process Y are $R_Y(\tau)$ and $C_Y(\tau)$, respectively. In the following theorem we prove the relationship between the auto-correlation and covariance functions of processes Y and X. For that, we need to define the distribution of all differences between two batches of samples as follows.

Definition 1 (Distribution between two batches of samples)**.** *Consider a RB of size N. Consider taking two different random permutations (batches), denoted by B_a and B_b, both of length K in two possibly different time points, t_a and $t_b = t_a + \tau$. Let τ'_K be a random variable where its distribution and expectation are denoted by $F_{\tau'_K}(\cdot)$ and $E_{\tau'_K}(\cdot)$, is the probability of each difference between each sample of B_a and B_b.*

We note that the support of τ' is $\tau - N + 1 \leq \tau'_K \leq \tau + N - 1$. Then,

Theorem 1 (Auto-Correlation and Covariance)**.** *Let τ be the difference between two time steps of the processes Y.*

Then:

$$R_Y(\tau) = \mathbb{E}_{\tau'_K}\left[R_Z(\tau'_K)\right], \\ C_Y(\tau) = \mathbb{E}_{\tau'_K}\left[C_Z(\tau'_K)\right]. \tag{1}$$

The proof for Theorem 1 is in Section B.2 in the supplementary material. We note two things. First, we note that we did not specify how the sampling is done from the RB and it is expressed by the random variable τ' from Definition 1, i.e., Eq. (1) is a general expression. Second, we note that we express the correlation using process Z and not process X directly, but process Z auto-correlation and covariance can be computed directly in any practical case using the relation $Z_t = f(X_t)$.

For the specific case of "unordered sampling without replacement", we express the relation between the second moments of Z and Y explicitly through the distribution of τ'.

Lemma 4 (Distribution for uniform batches)**.** *The random variable τ' distribution for "unordered sampling without replacement" is*

$$P(\tau') = \begin{cases} \frac{N-|d|}{N^2} & \tau' = \tau + d, \\ & d \in \{-N+1, \dots 0, \dots N-1\} \\ 0 & \tau' < \tau - N + 1 \text{ or } \tau' > \tau + N - 1 \end{cases}.$$

The proof for Lemma 4 is in the Section B.3 in the supplementary material. In the following corollary we state the exact dependence in the case of random sampling of K samples from a RB with size N.

Corollary 1. *Consider process Z where sampling is according to "unordered sampling without replacement". Then, the auto-correlation and covarinace of the process Y are:*

$$R_Y(\tau) = \frac{1}{N^2}\sum_{d=-N+1}^{N-1}(N-|d|)R_Z(d+\tau)$$

$$C_Y(\tau) = \frac{1}{N^2}\sum_{d=-N+1}^{N-1}(N-|d|)C_Z(d+\tau)$$

The proof for Corollary 1 is in the Section B.4 in the supplementary material. We see that using a RB reduces the autocorrelation and covaraince of process Z by factor of N. Interestingly, this reduction is independent of K. This result proves the de-correlation effect of using RBs and provides an explicit expression for that.

4. Replay Buffers in Reinforcement Learning

In the previous section we analyzed properties of stochastic processes that go through a RB. In this section we analyze RBs in RL. Stabilizing learning in modern off-policy deep RL algorithms, such as Deep Q-Networks (Mnih et al., 2013) or DDPG (Lillicrap et al., 2015), is based on saving past observed transitions in a RB. Even though its use is extensive, the theoretical understanding of sampling batches mechanism from a RB is still quite limited. This is our focus in this section.

We begin with describing the setup that will serve us in this section. Then we connect between the random processes as defined in Section 3 and common stochastic updates used in RL. We then describe an RB-based actor-critic algorithm that uses a batch of K samples from the RB in each parameters update step. This type of algorithm serves as a basic example for popular usages of RBs in RL. We note that other versions of RB-based RL algorithms (such as deep RL algorithms, value-based algorithms, discounted settings of the value function) can be analyzed with the stochastic processes tools we provide in this work. Finally, we present a full convergence proof for the RB-based actor critic algorithm.

Despite its popularity, to the best of our knowledge, there is only handful of proofs that consider RB in RL algorithm analysis (e.g., Di-Castro Shashua et al., 2021 or Lazic et al., 2021). Most of the convergence proofs for off-policy algorithms assume that a single sample is sampled from the RB. Di-Castro Shashua et al. (2021) proved for the first time the convergence of an RB-based algorithm. However, their algorithm and technical tools were focused on the sim-to-real challenge with multiple MDP environments, and they focused only on a single sample batch from the RB instead of K (which complicates the proof). Therefore, for completeness and focusing on the RB properties, we provide a proof for RB-based algorithms, with a single MDP environment and a batch of K samples. Similarly to previous works, we consider here a setup with linear function approximation (Bertsekas & Tsitsiklis, 1996).

4.1. Setup for Markov Decision Process

An environment in RL is modeled as a Markov Decision Process (MDP; Puterman, 1994), where $\mathcal{S}$ and $\mathcal{A}$ are the state space and action space, respectively. We let $P(S'|S,A)$ denote the probability of transitioning from state $S \in \mathcal{S}$ to state $S' \in \mathcal{S}$ when applying action $A \in \mathcal{A}$. We consider a probabilistic policy $\pi_\theta(A|S)$, parameterized by $\theta \in \Theta \subset \mathbb{R}^d$ which expresses the probability of the agent to choose an action A given that it is in state S. The MDP measure $P(S'|S,A)$ and the policy measure $\pi_\theta(A|S)$ induce together a Markov Chain (MC) measure $P_\theta(S'|S)$. We let μ_θ denote the stationary distribution induced by the policy π_θ. The reward function is denoted by $r(S,A)$.

In this work we focus on the *average reward* setting [5]. The

[5]The discount factor settings can be obtained in similar way to

goal of the agent is to find a policy that maximizes the average reward that the agent receives during its interaction with the environment. Under an ergodicity assumption, the average reward over time eventually converges to the expected reward under the stationary distribution (Bertsekas, 2005):

$$\eta_\theta \triangleq \lim_{T\to\infty} \frac{\sum_{t=0}^{T} r(S_t, A_t)}{T} = \mathbb{E}_{S\sim\mu_\theta, A\sim\pi_\theta}[r(S, A)]. \tag{2}$$

The state-value function evaluates the overall expected accumulated rewards given a starting state S and a policy π_θ

$$V^{\pi_\theta}(S) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty}(r(S_t, A_t) - \eta_\theta)\middle| S_0 = S, \pi_\theta\right], \tag{3}$$

where the actions follow the policy $A_t \sim \pi_\theta(\cdot|S_t)$ and the next state follows the transition probability $S_{t+1} \sim P(\cdot|S_t, A_t)$.

Let $O = \{S, A, S'\}$ be a transition from the environment. Let O_t be a transition at time t. The temporal difference error $\delta(O)$ (TD; Bertsekas & Tsitsiklis, 1996) is a random variable based on a single sampled transition from the RB,

$$\delta(O) = r(S, A) - \eta + \phi(S')^\top w - \phi(S)^\top w, \tag{4}$$

where $\hat{V}_w^{\pi_\theta}(S) = \phi(S)^\top w$ is a linear approximation for $V^{\pi_\theta}(S)$, $\phi(S) \in \mathbb{R}^d$ is a feature vector for state S and $w \in \mathbb{R}^d$ is the critic parameter vector.

4.2. Replay Buffer as a Random Process in RL

In Section 3 we compared between properties of general random process X going through a RB and yielding a process Y. In the RL context we have $X_t \triangleq O_t$, meaning our basic component is a single transition of state-action-next-state observed at time t. In addition, we defined $Z_t \triangleq f(X_t)$ process where $f(\cdot)$ is a general function. In RL, $f(\cdot)$ is commonly defined as the value function, the Q-function, the TD-error, the empirical average reward, the critic or actor gradients or any other function that computes a desirable update, based on an observed transition O. Common RL algorithms that use a single $f(O_t)$ computation in the parameters update step are commonly referred as *on-policy* algorithms where they update their parameters based only on the last observed transition in the Markov chain. See Figure 1 for a comparison between on-line updates and RB-based updates. Using the formulation of random processes we presented in Section 3, we can characterize the on-line

current setup.

Algorithm 1 Linear Actor Critic with RB samples

1: Initialize Replay Buffer RB with size N.
2: Initialize actor parameters θ_0, critic parameters w_0 and average reward estimator η_0.
3: Learning steps $\{\alpha_t^\eta\}$, $\{\alpha_t^w\}$, $\{\alpha_t^\theta\}$.
4: **for** $t = 0, \ldots$ **do**
5: Interact with MDP M according to policy π_{θ_t} and add the transition $\{S_t, A_t, r(S_t, A_t), S_{t+1}\}$ to RB$_t$.
6: Sample J_t - K random time indices form RB_t. Denote the corresponding transitions as $\{O_j\}_{j\in J_t}$.
7: $\delta(O_j) = r(S_j, A_j) - \eta_t + \phi(S'_j)^\top w_t - \phi(S_j)^\top w_t$
8: Update average reward $\eta_{t+1} = \eta_t + \alpha_t^\eta(\frac{1}{K}\sum_{j\in J_t} r(S_j, A_j) - \eta_t)$
9: Update critic $w_{t+1} = w_t + \alpha_t^w \frac{1}{K}\sum_{j\in J_t}\delta(O_j)\phi(S_j)$
10: Update actor $\theta_{t+1} = \Gamma\big(\theta_t - \alpha_t^\theta \frac{1}{K}\sum_{j\in J_t}\delta(O_j)\nabla_\theta \log\pi_\theta(A_j|S_j)\big)$
11: **end for**

updates, based on a single last transition as follows:

$$\begin{aligned} Z_t^{\text{reward}} &= f_{\text{reward}}(O_t) = r(S_t, A_t) - \eta_t \\ Z_t^{\text{critic}} &= f_{\text{critic}}(O_t) = \delta(O_t)\phi(S_t) \\ Z_t^{\text{actor}} &= f_{\text{actor}}(O_t) = \delta(O_t)\nabla\log\pi_\theta(A_t|S_t) \end{aligned}$$

When using RB-based off-policy algorithms, the parameters updates are computed over an average of K functions which are based on K transitions that were sampled randomly from the last stored N transitions. This exactly matches our definition of the process Y: $Y_t = \frac{1}{K}\sum_{j\in J_t} f(X_j) = \frac{1}{K}\sum_{j\in J_t} Z_j$. The following updates are typical in RB-based off-policy algorithms:

$$\begin{aligned} Y_t^{\text{reward}} &= \frac{1}{K}\sum_{j\in J_t} Z_j^{\text{reward}} = \frac{1}{K}\sum_{j\in J_t} r(S_j, A_j) - \eta_t \\ Y_t^{\text{critic}} &= \frac{1}{K}\sum_{j\in J_t} Z_j^{\text{critic}} = \frac{1}{K}\sum_{j\in J_t}\delta(O_j)\phi(S_j) \\ Y_t^{\text{actor}} &= \frac{1}{K}\sum_{j\in J_t} Z_j^{\text{actor}} = \frac{1}{K}\sum_{j\in J_t}\delta(O_j)\nabla\log\pi_\theta(A_j|S_j) \end{aligned} \tag{5}$$

In Algorithm 1, we present a linear actor critic algorithm based on RB samples where the algorithm updates the actor and critic using a random batch of transitions from the RB. In Section 4.5 we show how the results from Section 3 regarding a random process that is pushed into the RB, and the definitions of X and Y processes are helpful in proving the asymptotic convergence of this algorithm.

4.3. Linear Actor Critic with RB samples Algorithm

The basic RB-based algorithm we analyze in this work is presented in Algorithm 1. We propose a two time scale

linear actor critic optimization scheme (similarly to Konda & Tsitsiklis, 2000), which is an RB-based version of Bhatnagar et al. (2008) algorithm. Our algorithm is fully described by $W_t = [RB_t, J_t]$ and by the algorithm updates Y_t^{reward}, Y_t^{critic} and Y_t^{actor} described in equation (5). See Figure 2 for a visualized flow diagram of Algorithm 1.

In algorithm 1 we consider an environment, modeled as an MDP M, and we maintain a replay buffer RB with capacity N. The agent collects transitions $\{S, A, r(S, A), S'\}$ from the environment and stores them in the RB. We train the agent in an off-policy manner. At each time step t, the agent samples J_t – a subset of K random time indices from RB_t which defines the random transitions batch for optimizing the average reward, critic and actor parameters. Note that for the actor updates, we use a projection $\Gamma(\cdot)$ that projects any $\theta \in \mathbb{R}^d$ to a compact set Θ whenever $\theta \notin \Theta$.

4.4. Expectations of critic and actor updates in Algorithm 1

We divide the convergence analysis of Algorithm 1 into two parts. The first part, presented in this section, is unique to our paper - we describe in Lemmas 5 and 6 a closed form of the expectations of the actor and critic updates, based on a random batch of K transitions from the RB. In the second part, presented in Section 4.5, we use Stochastic Approximation (SA) tools for proving the algorithm updates convergence, based on the results from Lemmas 5 and 6. We note that Section 4.5 follows the steps of the convergence proofs presented by Di-Castro Shashua et al. (2021) and Bhatnagar et al. (2009).

For time $t - n + 1$ where $1 \leq n \leq N$, we define the induced MC with a corresponding policy parameter θ_{t-n+1}. For this parameter, we denote the corresponding state distribution vector ρ_{t-n+1} and a transition matrix P_{t-n+1} (both induced by the policy $\pi_{\theta_{t-n+1}}$). Finally, we define the following diagonal matrix $D_{t-n+1} \triangleq \text{diag}(\rho_{t-n+1})$ and the reward vector r_{t-n+1} with elements $r_{t-n+1}(S) = \sum_A \pi_{\theta_{t-n+1}}(A|S) r(S, A)$. Based on these definitions we define

$$\begin{aligned} C_t &\triangleq \frac{1}{N} \sum_{n=1}^{N} D_{t-n+1} \left(P_{t-n+1} - I\right) \\ b_t &\triangleq \frac{1}{N} \sum_{n=1}^{N} D_{t-n+1} \left(r_{t-n+1} - \eta_\theta e\right). \end{aligned} \tag{6}$$

where I is the identity matrix and e is a vector of ones. Let $D_\theta \triangleq \text{diag}(\mu_\theta)$ and define

$$C_\theta \triangleq D_\theta \left(P_\theta - I\right), \quad b_\theta \triangleq D_\theta \left(r_\theta - \eta_\theta e\right). \tag{7}$$

In our RB setting, since we have at time t a RB with the last N samples, C_t and b_t in Equation (6) represent a superposition of all related elements of these samples. For proving the convergence of the critic, we assume the policy is fixed. Then, when $t \to \infty$, $\rho_{t-n+1} \to \mu_\theta$ for all index n. This means that the induced MC is one for all the samples in the RB, so the sum over N disappear for C_θ and b_θ.

The following two lemmas compute the expectation of the critic and actor updates when using a random batch of K samples. The expectations are over all possible random batches sampled from the RB. Recall that $\bar{J}_t \subset \{1, \ldots, N\}$ and $\mathbb{C}_{N,K}$ is the set of all possible subsets $\bar{J}$ for specific N and K. These lemmas are the main results for proving convergence of RB-based RL algorithms.

Lemma 5. *Assume we have a RB with N transitions and we sample random K transitions from the RB. Then:*

$$\mathbb{E}_{\bar{J}_t \sim \mathbb{C}_{N,K}, \{O_{t-n+1}\}_{n \in \bar{J}_t} \sim RB_t} \left[\frac{1}{K} \sum_{n \in \bar{J}_t} \delta(O_{t-n+1}) \phi(S_{t-n+1}) \right] = \Phi^\top C_\theta \Phi w + \Phi^\top b_\theta,$$

where C_θ and b_θ are defined in (7).

Lemma 6. *Assume we have a RB with N transitions and we sample random K transitions from the RB. Then:*

$$\begin{aligned} &\mathbb{E}_{\bar{J}_t \sim \mathbb{C}_{N,K}, \{O_{t-n+1}\}_{n \in \bar{J}_t} \sim RB_t} \\ &\left[\frac{1}{K} \sum_{n \in \bar{J}_t} \delta^{\pi_\theta}(O_{t-n+1}) \nabla_\theta \log \pi_\theta(A_{t-n+1}|S_{t-n+1}) \right] \\ &= \nabla_\theta \eta_\theta - \sum_S \mu_\theta(S) \Big(\phi(S)^\top \nabla_\theta w^{\pi_\theta} - \nabla_\theta \bar{V}^{\pi_\theta}(S) \Big), \end{aligned}$$

where $\bar{V}^{\pi_\theta}(S) = \sum_{A \in \mathcal{A}} \pi_\theta(A|S) \big(r(S, A) - \eta_\theta + \sum_{S' \in \mathcal{S}} P(S'|S, A) \phi(S')^\top w^{\pi_\theta} \big)$.

The proofs for Lemmas 5 and 6 are in sections C.1 and D.1, respectively, in the supplementary material.

4.5. Asymptotic Convergence of Algorithm 1

We are now ready to present the convergence theorems for the critic and actor in Algorithm 1. In the proof of our theorems we use tools from Stochastic Approximation (SA) (Kushner & Yin, 2003; Borkar, 2009; Bertsekas & Tsitsiklis, 1996), a standard tool in the literature for analyzing iterations of processes such as two time scale Actor-Critic in the context of RL.

We showed in Lemma 2 that the process $W_t = [RB_t, J_t]$ of sampling K random transitions from the RB is a Markov process. In addition, we showed in Lemma 3 that if the original Markov chain is irreducible and aperiodic, then also the RB Markov process is irreducible and aperiodic. This property is required for the existence of unique stationary distribution and for proving the convergence of the iterations

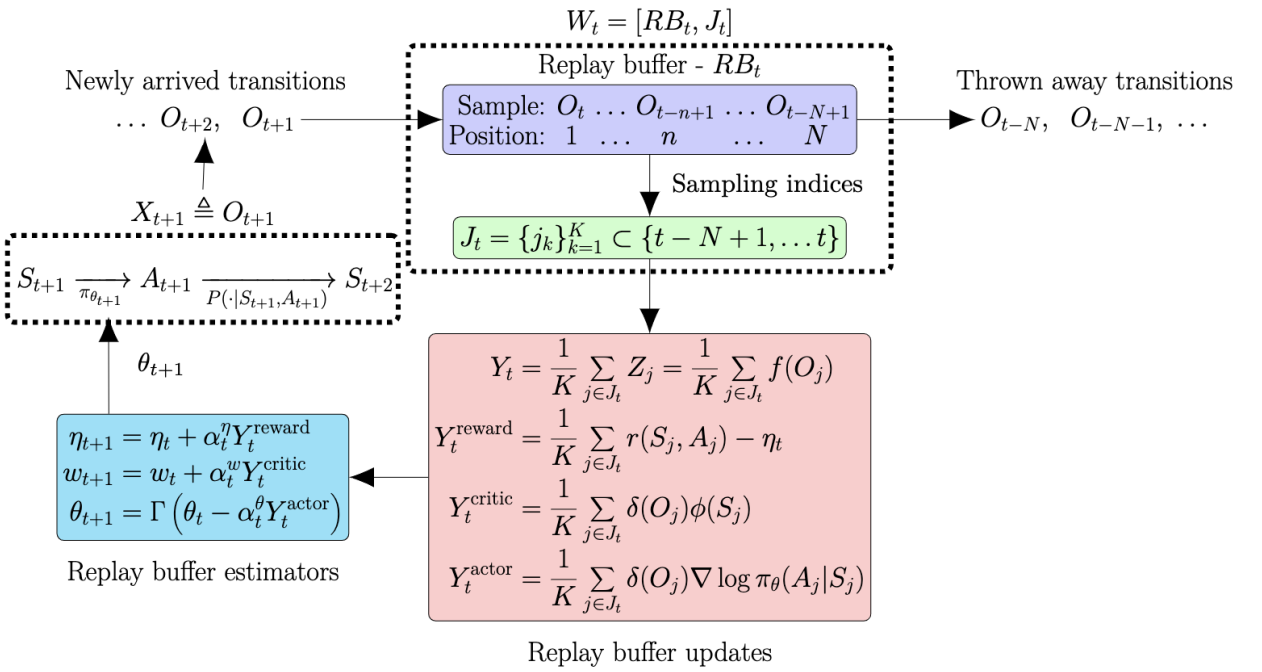

Figure 2. Replay buffer in reinforcement learning flow diagram: The random processes described in Figure 1 are reflected in Algorithm 1. Here the random process that enters the RB is O which is a tuple of (S, A, S'). The RB stores the last N transitions $\{O_t, \dots O_{t-N-1}\}$ in positions $(1, \dots, N)$, respectively. As time proceeds and $t > N$, old transition are thrown away from the RB. At each time step t, a random subset of K time steps is sampled from the RB and is denoted as J_t. W is simply $[RB, J]$. In Algorithm 1 we have three different updates, Y_t^{reward}, Y_t^{critic} and Y_t^{actor}, all are averages over functions of transitions sampled from the RB. Then the parameters are updated accordingly. Finally, the policy parameter θ_{t+1} is used to sample the action in transition O_{t+1} that later enters to the RB.

in Algorithm 1 using SA tools. We note that proving convergence for a general function approximation is hard. We choose to demonstrate the convergence for a linear function approximation (LFA; Bertsekas & Tsitsiklis, 1996), in order to keep the convergence proof as simple as possible while focusing in the proof on the RB and random batches aspects of the algorithm.

We present several assumptions that are necessary for proving the convergence of Algorithm 1. Assumption 4 is needed for the uniqueness of the convergence point of the critic. In addition, we choose a state S^* to be of value 0, i.e., $V^{\pi_\theta}(S^*) = 0$ (due to Assumption 2, S^* can be any of $S \in \mathcal{S}$). Assumption 5 is required in order to get a *with probability 1* using the SA convergence. In our actor-critic setup we need two time-scales convergence, thus, in this assumption the critic is a 'faster' recursion than the actor.

Assumption 1. *1. The set Θ is compact. 2. The reward $|r(\cdot,\cdot)| \leq 1$ for all $S \in \mathcal{S}, A \in \mathcal{A}$.*

Assumption 2. *For any policy π_θ, the induced Markov chain of the MDP process $\{S_t\}_{t\geq 0}$ is irreducible and aperiodic.*

Assumption 3. *For any state–action pair (S, A), $\pi_\theta(A|S)$ is continuously differentiable in the parameter θ.*

Assumption 4. *1. The matrix Φ has full rank. 2. The functions $\phi(S)$ are Liphschitz in s and bounded. 3. For every $w \in \mathbb{R}^d$, $\Phi w \neq e$ where e is a vector of ones.*

Assumption 5. *The step-sizes $\{\alpha_t^\eta\}$, $\{\alpha_t^w\}$, $\{\alpha_t^\theta\}$, $t \geq 0$ satisfy $\sum_t^\infty \alpha_t^\eta = \sum_t^\infty \alpha_t^w = \sum_t^\infty \alpha_t^\theta = \infty$, $\sum_t^\infty (\alpha_t^\eta)^2$, $\sum_t^\infty (\alpha_t^w)^2$, $\sum_t^\infty (\alpha_t^\theta)^2 < \infty$ and $\alpha_t^\theta = o(\alpha_t^w)$.*

Now we are ready to prove the following theorems, regarding Algorithm 1. We note that Theorem 2 and 3 state the critic and actor convergence.

Theorem 2. *(Convergence of the Critic to a fixed point) Under Assumptions 1-5, for any given π and $\{\eta_t\}, \{w_t\}$ as in the updates in Algorithm 1, we have $\eta_t \to \eta_\theta$ and $w_t \to w^\pi$ with probability 1, where w^π is obtained as a unique solution to $\Phi^\top C_\theta \Phi w + \Phi^\top b_\theta = 0$.*

The proof for Theorem 2 is in Section C in the supplementary material. It follows the proof for Lemma 5 in Bhatnagar et al. (2009). For establishing the convergence of the actor updates, we define additional terms. Let $\mathcal{Z}$ denote the set of asymptotically stable equilibria of the ODE $\dot{\theta} = \hat{\Gamma}(-\nabla_\theta \eta_\theta)$ and let $\mathcal{Z}^\epsilon$ be the ϵ-neighborhood of $\mathcal{Z}$. We define $\xi^{\pi_\theta} = \sum_S \mu_\theta(S)\Big(\phi(S)^\top \nabla_\theta w^{\pi_\theta} - \nabla_\theta \bar{V}^{\pi_\theta}(S)\Big)$.

Theorem 3. *(Convergence of the actor)*
Under Assumptions 1-5, given $\epsilon > 0$, $\exists \delta > 0$ *such that for* θ_t, $t \geq 0$ *obtained using Algorithm 1, if* $\sup_{\theta_t} \|\xi^{\pi_{\theta_t}}\| < \delta$, *then* $\theta_t \to \mathcal{Z}^\epsilon$ *as* $t \to \infty$ *with probability one.*

The proof for Theorem 3 is in Section D in the supplementary material. It follows the proof for Theorem 2 in Bhatnagar et al. (2009).

5. Related Work

Actor critic algorithms analysis: The convergence analysis of our proposed RB-based actor critic algorithm is based on the Stochastic Approximation method (Kushner & Clark, 2012). Konda & Tsitsiklis (2000) proposed the actor-critic algorithm, and established the asymptotic convergence for the two time-scale actor-critic, with TD(λ) learning-based critic. Bhatnagar et al. (2009) proved the convergence result for the original actor-critic and natural actor-critic methods. Di Castro & Meir (2010) proposed a single time-scale actor-critic algorithm and proved its convergence. Works on finite sample analysis for actor critic algorithms (Wu et al., 2020; Zou et al., 2019; Dalal et al., 2018) analyze the case of last transition update and do not analyze the RB aspects in these algorithms.

Recently, Di-Castro Shashua et al. (2021) proved for the first time the convergence of an RB-based actor critic algorithm. However, their algorithm and technical tools were focused on the sim-to-real challenge with multiple MDP environments, and they focused only on a single sample batch from the RB instead of K (which complicates the proof). We provide a proof for RB-based algorithms, with a single MDP environment and a batch of K samples.

Replay Buffer analysis: Experience replay (Lin, 1993) is a central concept for achieving good performance in deep reinforcement learning. Deep RB-based algorithms such as deep Q-learning (DQN, Mnih et al., 2013), deep deterministic policy gradient (DDPG; Lillicrap et al., 2015), actor critic with experience replay (ACER; Wang et al., 2016), Twin Delayed Deep Deterministic policy gradient (TD3, Fujimoto et al., 2018), Soft Actor Critic (SAC, Haarnoja et al., 2018) and many others use RBs to improve performance and data efficiency.

We focus mainly on works that provide some RB properties analysis. Zhang & Sutton (2017) and Liu & Zou (2018) study the effect of replay buffer size on the agent performance . Fedus et al. (2020) investigated through simulated experiments how the learning process is affected by the ratio of learning updates to experience collected. Other works focus on methods to prioritize samples in the RB and provide experimental results to emphasis performance improvement when using prioritized sampling from RB (Schaul et al., 2015; Pan et al., 2018; Zha et al., 2019; Horgan et al., 2018; Lahire et al., 2021). We, on the other hand, focus on the theoretical aspects of RB properties and provide convergence results for RB-based algorithms. Lazic et al. (2021) proposed a RB version for a regularized policy iteration algorithm. They provide an additional motivation for using RBs, in addition to the advantage of reduced temporal correlations: They claim that using RB in online learning in MDPs can approximate well the average of past value functions. Their analysis also suggests a new objective for sub-sampling or priority-sampling transitions in the RB, which differs priority-sampling objectives of previous work (Schaul et al., 2015).

Regarding RB analysis in Deep RL algorithms, Fan et al. (2020) performed a finite sample analysis on DQN algorithm (Mnih et al., 2013). In their analysis, they simplified the technique of RB with an independence assumption and they replaced the distribution over random batches with a fixed distribution. These assumptions essentially reduce DQN to the neural fitted Q-iteration (FQI) algorithm (Riedmiller, 2005). In our work we focus on asymptotic convergence and analyze explicitly the distribution of random batches from the RB.

6. Conclusions

In this work we analyzed RB and showed some basic random processes properties of it as ergodicity, stationarity, Markovity, correlation, and covariance. The latter two are of most interest since they can explain the success of modern RL algorithm based on RB. In addition, we showed quantitatively the relations between the RB size, batch size, and other factors.

In addition, we developed theoretical tools of stochastic process analysis for replay buffers. We provided an example of how to use these tools to analyze the convergence of an RB-based actor critic algorithm. Similarly, other common RB-based algorithms in reinforcement learning such as DQN (Mnih et al., 2013), DDPG (Lillicrap et al., 2015), TD3 (Fujimoto et al., 2018), SAC (Haarnoja et al., 2018) and many others can be analyzed now, using the tools we developed in this work.

As a future research, we propose two directions that are of great interest and complete the analysis we provided in this work:

1. **Spectrum analysis of the learning processes.** Since we adopted an approach of "Signals and Systems" with random signals (Oppenheim et al., 1997; Porat, 2008), one can use spectrum analysis in order to discover instabilities or cycles in the learning process.

2. **More complex RBs.** There is a large experimental body of work that tries to propose different schemes

of RBs. Some of them apply different independent on RL quantities sampling techniques while other apply dependent on RL quantities schemes (e.g., prioritized RB depends on the TD signal; Schaul et al., 2015). In this work we paved the first steps to apply analysis on such schemes (both dependent and independent).

References

Bertsekas, D. *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 2005.

Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-dynamic programming*. Athena Scientific, 1996.

Bhatnagar, S. and Kumar, S. A simultaneous perturbation stochastic approximation-based actor-critic algorithm for markov decision processes. *IEEE Transactions on Automatic Control*, 49(4):592–598, 2004.

Bhatnagar, S., Ghavamzadeh, M., Lee, M., and Sutton, R. S. Incremental natural actor-critic algorithms. In *Advances in neural information processing systems*, pp. 105–112, 2008.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. Natural actor–critic algorithms. *Automatica*, 45(11): 2471–2482, 2009.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Borkar, V. S. and Meyn, S. P. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38 (2):447–469, 2000.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. Finite sample analyses for td (0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Di Castro, D. and Meir, R. A convergent online single time scale actor critic algorithm. *The Journal of Machine Learning Research*, 11:367–410, 2010.

Di-Castro Shashua, S., Di Castro, D., and Mannor, S. Sim and real: Better together. *Advances in Neural Information Processing Systems*, 34, 2021.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Fedus, W., Ramachandran, P., Agarwal, R., Bengio, Y., Larochelle, H., Rowland, M., and Dabney, W. Revisiting fundamentals of experience replay. In *International Conference on Machine Learning*, pp. 3061–3071. PMLR, 2020.

Fujimoto, S., Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pp. 1587–1596. PMLR, 2018.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Horgan, D., Quan, J., Budden, D., Barth-Maron, G., Hessel, M., Van Hasselt, H., and Silver, D. Distributed prioritized experience replay. *arXiv preprint arXiv:1803.00933*, 2018.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Advances in neural information processing systems*, pp. 1008–1014. Citeseer, 2000.

Kushner, H. and Yin, G. G. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Kushner, H. J. and Clark, D. S. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media, 2012.

Laguna, M. and Marklund, J. *Business process modeling, simulation, and design*. Taylor & Francis, 2013.

Lahire, T., Geist, M., and Rachelson, E. Large batch experience replay. *arXiv preprint arXiv:2110.01528*, 2021.

Lazic, N., Yin, D., Abbasi-Yadkori, Y., and Szepesvari, C. Improved regret bound and experience replay in regularized policy iteration. *arXiv preprint arXiv:2102.12611*, 2021.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Lin, L.-J. Reinforcement learning for robots using neural networks. Technical report, Carnegie-Mellon Univ Pittsburgh PA School of Computer Science, 1993.

Liu, R. and Zou, J. The effects of memory replay in reinforcement learning. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 478–485. IEEE, 2018.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Norris, J. R. *Markov chains*. Number 2. Cambridge university press, 1998.

Oppenheim, A. V., Willsky, A. S., Nawab, S. H., Hernández, G. M., et al. *Signals & systems*. Pearson Educación, 1997.

Pan, Y., Zaheer, M., White, A., Patterson, A., and White, M. Organizing experience: a deeper look at replay mechanisms for sample-based planning in continuous state domains. *arXiv preprint arXiv:1806.04624*, 2018.

Porat, B. *Digital processing of random signals: theory and methods*. Courier Dover Publications, 2008.

Puterman, M. L. *Markov Decision Processes*. Wiley and Sons, 1994.

Riedmiller, M. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European conference on machine learning*, pp. 317–328. Springer, 2005.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.

Wu, Y., Zhang, W., Xu, P., and Gu, Q. A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*, 2020.

Zha, D., Lai, K.-H., Zhou, K., and Hu, X. Experience replay optimization. *arXiv preprint arXiv:1906.08387*, 2019.

Zhang, S. and Sutton, R. S. A deeper look at experience replay. *arXiv preprint arXiv:1712.01275*, 2017.

Zou, S., Xu, T., and Liang, Y. Finite-sample analysis for sarsa with linear function approximation. *arXiv preprint arXiv:1902.02234*, 2019.

A. Proofs for Lemmas in Section 3

A.1. Proof of Lemma 1

Proof. **Stationarity of (RB_t):** Recall that stationarity (in the strong sense) means that for $m = 1, 2, \ldots$, there are times $(t_1, t_2, \ldots, t_m)$ such that for all $\tau \in \mathbb{Z}$

$$F_X(X_{t_1+\tau}, \ldots, X_{t_m+\tau}) = F_X(X_{t_1}, \ldots, X_{t_m}),$$

where $F_X(X_{t_1}, \ldots, X_{t_m})$ is the cumulative distribution. Then,

$$\begin{aligned}
F_{RB}(RB_{t_1+\tau}, \ldots, RB_{t_m+\tau}) &\overset{(1)}{=} F_X(X_{t_1+\tau-N+1}, \ldots, X_{t_1+\tau}, \\
&\quad X_{t_2+\tau-N+1}, \ldots, X_{t_2+\tau}, \\
&\quad \ldots, \\
&\quad X_{t_m+\tau-N+1}, \ldots, X_{t_m+\tau}), \\
&\overset{(2)}{=} F_X(X_{t_1-N+1}, \ldots, X_{t_1}, \\
&\quad X_{t_2-N+1}, \ldots, X_{t_2}, \\
&\quad \ldots, \\
&\quad X_{t_m-N+1}, \ldots, X_{t_m}) \\
&\overset{(3)}{=} F_{RB}(RB_{t_1-N+1}, \ldots, RB_{t_1}),
\end{aligned}$$

where we use the the RB definition in (1), stationarity of X in (2), and expressing RB based on X in (3).

Stationarity of Y_t: Similarly, for $m = 1, 2, \ldots$, there are times $(t_1, t_2, \ldots, t_m)$ and we have

$$\begin{aligned}
F_Y(Y_{t_1+\tau}, \ldots, Y_{t_m+\tau}) &\overset{(1)}{=} F_X\left(\frac{1}{K}\sum_{j\in J_{t_1+\tau}} f(X_j), \ldots, \frac{1}{K}\sum_{j\in J_{t_m+\tau}} f(X_j)\right) \\
&\overset{(2)}{=} \sum_{J_{t_1+\tau},\ldots,J_{t_m+\tau}} F_{J_{t_1+\tau},\ldots,J_{t_m+\tau}}(j_1, \ldots, j_m)\times \\
&\quad F_X\left(\frac{1}{K}\sum_{j\in J_{t_1+\tau}} f(X_j), \ldots, \frac{1}{K}\sum_{j\in J_{t_m+\tau}} f(X_j)\middle| j_1, \ldots, j_m\right) \\
&\overset{(3)}{=} \sum_{J_{t_1},\ldots,J_{t_m}} F_{J_{t_1},\ldots,J_{t_m}}(j_1, \ldots, j_m)\times \\
&\quad F_X\left(\frac{1}{K}\sum_{j\in J_{t_1}} f(X_j), \ldots, \frac{1}{K}\sum_{j\in J_{t_m}} f(X_j)\middle| j_1, \ldots, j_m\right) \\
&\overset{(4)}{=} F_X\left(\frac{1}{K}\sum_{j\in J_{t_1}} f(X_j), \ldots, \frac{1}{K}\sum_{j\in J_{t_m}} f(X_j)\right) \\
&\overset{(5)}{=} F_Y(Y_{t_1}, \ldots, Y_{t_m}),
\end{aligned}$$

where in (1) we use the process Y definition, in (2) we use iterated expectation, in (3) we use both the stationarity of X and J, in (4) we use again iterated expectation, and in (5) we use Y definition.

□

A.2. Proof of Lemma 2

Proof. Proving Markovity requires that

$$P(RB_{t+1}|RB_t, RB_{t-1}, \ldots, RB_0) = P(RB_{t+1}|RB_t). \tag{A.1}$$
$$P(W_{t+1}|W_t, W_{t-1}, \ldots, W_0) = P(W_{t+1}|W_t). \tag{A.2}$$

We start with proving the Markovity of RB_t. Let us denote $X_{n_1}^{n_2}(t) \triangleq \{X_{t-n_2+1}, \ldots, X_{t-n_1+1}|RB_t\}$ as the set of random variables from process X stored in the RB at time t in positions n_1 to n_2. Note that when a new transition is pushed to the RB into position $n = 1$, the oldest transition in position $n = N$ is thrown away, and all the transitions in the RB move one index forward. We present here some observations regarding the RB that will help us through the proof:

$$RB_t = X_1^N(t) = \{X_{t-N+1}, \ldots, X_{t-n+1} \ldots, X_t\} \quad \text{(RB definition).} \tag{A.3}$$
$$X_1^N(t+1) = \{X_{t+1}\} \cup X_1^{N-1}(t) \tag{A.4}$$
$$X_1^{N-1}(t) \subset X_1^N(t) \tag{A.5}$$
$$X_t \in X_1^N(t) \tag{A.6}$$
$$P(X_{t+1}|X_t, \ldots X_0) = P(X_{t+1}|X_t) \quad \text{(Since } X_t \text{ is assumed to be Markovian).} \tag{A.7}$$
$$P(a, b|c_1, c_2, \ldots) = P(a|b, c_1, c_2, \ldots) \cdot P(b|c_1, c_2, \ldots) \quad \text{(Expressing joint probability as a conditional probabilities product).} \tag{A.8}$$
$$P(a|b) = p(a) \quad \text{(If a and b are independent) .} \tag{A.9}$$

Computing the l.h.s. of equation (A.1) yields

$$\begin{aligned} P\left(RB_{t+1}|RB_t, \ldots, RB_0\right) &\overset{(A.3)}{=} P\left(X_1^N(t+1)\middle|X_1^N(t), \ldots, X_1^N(0)\right) \\ &\overset{(A.4)}{=} P\left(X_{t+1}, X_1^{N-1}(t)\middle|X_1^N(t), \ldots, X_1^N(0)\right) \\ &\overset{(A.8)}{=} P\left(X_{t+1}\middle|X_1^{N-1}(t), X_1^N(t), \ldots, X_1^N(0)\right) \cdot P\left(X_1^{N-1}(t)\middle|X_1^N(t), \ldots, X_1^N(0)\right) \\ &\overset{(A.5),(A.6),(A.7)}{=} P\left(X_{t+1}|X_t\right) \end{aligned}$$

Similarly, computing the r.h.s of (A.1) yields

$$\begin{aligned} P\left(RB_{t+1}|RB_t\right) &\overset{(A.3)}{=} P\left(X_1^N(t+1)\middle|X_1^N(t)\right) \\ &\overset{(A.4)}{=} P\left(X_{t+1}, X_1^{N-1}(t)\middle|X_1^N(t)\right) \\ &\overset{(A.8)}{=} P\left(X_{t+1}\middle|X_1^{N-1}(t), X_1^N(t)\right) \cdot P\left(X_1^{N-1}(t)\middle|X_1^N(t)\right) \\ &\overset{(A.5),(A.6),(A.7)}{=} P\left(X_{t+1}|X_t\right) \end{aligned}$$

Both sides of (A.1) are equal and therefore RB_t is Markovian. In addition we have that for $t \geq N$:

$$P\left(RB_{t+1}|RB_t\right) = \begin{cases} P\left(X_{t+1}|X_t\right) & \text{if } X_t \in X_1^1(t) \text{ and } RB_{t+1} = \{X_{t+1}\} \cup X_1^{N-1}(t), \\ 0 & \text{otherwise.} \end{cases}$$

Recall that W_t is defined as:

$$W_t = [RB_t, J_t] \tag{A.10}$$

where $J_t \subset \{t - N + 1, \ldots, t\}$ is a random subset of K time indices. By their definition, RB_t and J_t are independent for all t. Computing the l.h.s. of equation (A.2) yields

$$\begin{aligned} P\left(W_{t+1}|W_t, \ldots, W_0\right) &\overset{(A.10)}{=} P\left(RB_{t+1}, J_{t+1}|RB_t, J_t, \ldots, RB_0, J_0\right) \\ &\overset{(A.8)}{=} P\left(RB_{t+1}|J_{t+1}, RB_t, J_t, \ldots, RB_0, J_0\right) \cdot P\left(J_{t+1}|RB_t, J_t, \ldots, RB_0, J_0\right) \\ &\overset{(A.1),(A.9)}{=} P\left(RB_{t+1}|RB_t\right) \cdot P\left(J_{t+1}\right) \\ &\overset{(A.9)}{=} P\left(RB_{t+1}, J_{t+1}|RB_t, J_t\right) \\ &\overset{(A.10)}{=} P\left(W_{t+1}|W_t\right) \end{aligned}$$

We have the required result in (A.2), therefore W_t is Markovian.

In addition, If J_{t+1} is sampled according to "unordered sampling without replacement" (defined in Section 2.2), then for $t \geq N$:

$$P\left(W_{t+1}|W_t\right) = P\left(RB_{t+1}|RB_t\right) \cdot P\left(J_{t+1}\right) = \begin{cases} \frac{1}{\binom{N}{K}} P\left(X_{t+1}|X_t\right) & \text{if } X_t \in X_1^1(t) \text{ and } RB_{t+1} = \{X_{t+1}\} \cup X_1^{N-1}(t) \\ & \forall J_{t+1} \in \mathbb{C}_{N,K}, \\ 0 & \text{otherwise.} \end{cases}$$

□

A.3. Proof of Lemma 3

Proof. We prove by contradiction. Let us assume that the process RB is neither aperiodic nor irreducible. If it is aperiodic, then one of the indices in the RB is aperiodic. Without loss of generality, let us assume that this is the l delayed time-steps index. But since in this index we have aperiodic process, i.e., it is the process X delayed in l steps, it contradicts the assumption that X is aperiodic. We prove irreducibility in a similar way.

Since the process Y is a deterministic function of the process RB, it must be aperiodic and irreducible as well, otherwise it will contradict the aperiodicity and irreducibility of the process RB. Finally, since $f(\cdot)$ is deterministic function, and since for each t, Y_t is an image of an ergodic process X_t, i.e., each $x \in X_t$ is visited infinitely often, and as a results each point $y \in Y_t$ of the image of $f(\cdot)$ is visited infinitely often, otherwise, it contradicts the deteministic nature of $f(\cdot)$ or the ergodicity of X. □

B. Auto Correlation and Covariance proofs

B.1. Proof of Theorem 1

Proof. Let $\bar{J}_t \subset \{1, \dots N\}$ and $\bar{J}_{t+\tau} \subset \{1, \dots N\}$ be subsets of K indices each. We begin with calculating the auto-correlation of process Z_t.

$$\begin{aligned}
R_Y(\tau) &= \mathbb{E}[Y_t Y_{t+\tau}] \\
&\overset{(1)}{=} \mathbb{E}\left[\frac{1}{K}\sum_{n\in\bar{J}_t} Z_{t-n+1} \cdot \frac{1}{K}\sum_{m\in\bar{J}_{t+\tau}} Z_{t+\tau-m+1}\right] \\
&\overset{(2)}{=} \mathbb{E}\left[\frac{1}{K}\sum_{n\in\bar{J}_t} f(X_{t-n+1}) \cdot \frac{1}{K}\sum_{m\in\bar{J}_{t+\tau}} f(X_{t+\tau-m+1})\right] \\
&\overset{(3)}{=} \mathbb{E}_{\bar{J}_t,\bar{J}_{t+\tau}\sim\mathbb{C}_{N,K},\{X_{t-n+1}\}_{n\in\bar{J}_t}\sim RB_t,\{X_{t+\tau-m+1}\}_{n\in\bar{J}_{t+\tau}}\sim RB_{t+\tau}}\left[\frac{1}{K}\sum_{n\in\bar{J}_t} f(X_{t-n+1}) \cdot \frac{1}{K}\sum_{m\in\bar{J}_{t+\tau}} f(X_{t+\tau-m+1})\right] \\
&\overset{(4)}{=} \mathbb{E}_{\bar{J}_t,\bar{J}_{t+\tau}\sim\mathbb{C}_{N,K}}\left[\mathbb{E}_{\{X_{t-n+1}\}_{n\in\bar{J}_t}\sim RB_t,\{X_{t+\tau-m+1}\}_{n\in\bar{J}_{t+\tau}}\sim RB_{t+\tau}}\left[\frac{1}{K}\sum_{n\in\bar{J}_t} f(X_{t-n+1}) \cdot \frac{1}{K}\sum_{m\in\bar{J}_{t+\tau}} f(X_{t+\tau-m+1})\middle|\bar{J}_t,\bar{J}_{t+\tau}\right]\right] \\
&\overset{(5)}{=} \mathbb{E}_{\bar{J}_t,\bar{J}_{t+\tau}\sim\mathbb{C}_{N,K}}\left[\mathbb{E}_{\{X_{t-n+1}\}_{n\in\bar{J}_t}\sim RB_t,\{X_{t+\tau-m+1}\}_{n\in\bar{J}_{t+\tau}}\sim RB_{t+\tau}}\left[\mathbb{E}_{n\sim\bar{J}_t,m\sim\bar{J}_{t+\tau}}\left[f(X_{t-n+1})\cdot f(X_{t+\tau-m+1})\right]\middle|\bar{J}_t,\bar{J}_{t+\tau}\right]\right] \\
&\overset{(6)}{=} \mathbb{E}_{\bar{J}_t,\bar{J}_{t+\tau}\sim\mathbb{C}_{N,K}}\left[\mathbb{E}_{n\sim\bar{J}_t,m\sim\bar{J}_{t+\tau}}\left[\mathbb{E}_{X_{t-n+1},X_{t+\tau-m+1}}\left[f(X_{t-n+1})\cdot f(X_{t+\tau-m+1})\right]\middle|\bar{J}_t,\bar{J}_{t+\tau}\right]\right] \\
&\overset{(7)}{=} \mathbb{E}_{\bar{J}_t,\bar{J}_{t+\tau}\sim\mathbb{C}_{N,K}}\left[\mathbb{E}_{n\sim\bar{J}_t,m\sim\bar{J}_{t+\tau}}\left[\mathbb{E}\left[Z_{t-n+1}Z_{t+\tau-m+1}\right]\middle|\bar{J}_t,\bar{J}_{t+\tau}\right]\right] \\
&\overset{(8)}{=} \mathbb{E}_{\bar{J}_t,\bar{J}_{t+\tau}\sim\mathbb{C}_{N,K}}\left[\mathbb{E}_{n\sim\bar{J}_t,m\sim\bar{J}_{t+\tau}}\left[R_Z(\tau+n-m)\middle|\bar{J}_t,\bar{J}_{t+\tau}\right]\right] \\
&\overset{(9)}{=} \mathbb{E}_{\tau'\sim\tilde{J}_\tau}\left[R_Z(\tau')\right]
\end{aligned}$$

where in (1) we used the definition of Y using the indices subsets $\bar{J}_t$ and $\bar{J}_{t+\tau}$. In (2) used the definition of Z and in (3) we wrote the expectation explicitly. In (4) we used the conditional expectation and in (5) we wrote $\frac{1}{K}\sum_{n\in\bar{J}_t} f(\cdot)$ and $\frac{1}{K}\sum_{m\in\bar{J}_{t+\tau}} f(\cdot)$ as an expectations since given the subsets $\bar{J}_t$ and $\bar{J}_{t+\tau}$, the probability of sampling index n or m from the RB is uniform and equals $\frac{1}{K}$. In (6) we switched between the expectations, since given the subsets $\bar{J}_t$ and $\bar{J}_{t+\tau}$, the samples $\{X_{t-n_k+1}\}_{k=1}^K$ and $\{X_{t+\tau-m_k+1}\}_{k=1}^K$ are independent. In (7) we used again the definition of Z and in (8) we used the definition of the auto-correlation function of Z. In (9) we defined $\tau' = \tau + n - m$ to be the time difference between each couple of indices from $\bar{J}_t$ and $\bar{J}_{t+\tau}$. Note that $\tau' \in \tilde{J}_\tau$ where $\tilde{J}_\tau = \{\tau - N + 1, \dots, \tau + N - 1\}$.

The calculation for the covariance $C_Y(\tau)$ follows the same steps as we did for $R_Y(\tau)$. □

B.2. Alternative Proof of Theorem 1

Proof. Let $\bar{J}_t \subset \{1, \dots N\}$ and $\bar{J}_{t+\tau} \subset \{1, \dots N\}$ be subsets of K indices each. We begin with calculating the auto-correlation of process Y_t.

$$\begin{aligned}
R_Y(\tau) &= \mathbb{E}[Y_t Y_{t+\tau}] \\
&= \mathbb{E}\left[\frac{1}{K}\sum_{n\in\bar{J}_t} Z_{t-n+1} \cdot \frac{1}{K}\sum_{m\in\bar{J}_{t+\tau}} Z_{t+\tau-m+1}\right] \\
&= \mathbb{E}_{\bar{J}_t,\bar{J}_{t+\tau}\sim\mathbb{C}_{N,K}}\left[\mathbb{E}\left[\frac{1}{K}\sum_{n\in\bar{J}_t} Z_{t-n+1} \cdot \frac{1}{K}\sum_{m\in\bar{J}_{t+\tau}} Z_{t+\tau-m+1}\middle|\bar{J}_t, \bar{J}_{t+\tau}\right]\right] \\
&= \mathbb{E}_{\bar{J}_t,\bar{J}_{t+\tau}\sim\mathbb{C}_{N,K}}\left[\mathbb{E}\left[\frac{1}{K^2}\sum_{n\in\bar{J}_t}\sum_{m\in\bar{J}_{t+\tau}} Z_{t-n+1} Z_{t+\tau-m+1}\middle|\bar{J}_t, \bar{J}_{t+\tau}\right]\right].
\end{aligned}$$

Next, we will go from the definite times defined by $\bar{J}_t$ and $\bar{J}_{t+\tau}$ to all the time differences defined by the same $\bar{J}_t$ and $\bar{J}_{t+\tau}$. Let $D_t \triangleq \{d_i\}_{i=1}^K$ be based on $\bar{J}_t$ and be a set of difference times where d_1 is the time from the RB beginning to the first sample, d_2 is the time from the first sample to the second samples, and so on until d_K which is the time different between the one before the last sample to the last sample. Similarly, we define $D_{t+\tau} \triangleq \{d_j\}_{j=1}^K$ to be based on $\bar{J}_{t+\tau}$. Therefore,

$$\begin{aligned}
R_Y(\tau) &= \mathbb{E}_{\bar{J}_t,\bar{J}_{t+\tau}\sim\mathbb{C}_{N,K}}\left[\mathbb{E}\left[\frac{1}{K^2}\sum_{i=1}^K\sum_{j=1}^K Z_{t+\sum_{i'=1}^i d_{i'}} Z_{t+\tau+\sum_{j'=1}^j d_{j'}}\middle|\bar{J}_t, \bar{J}_{t+\tau}\right]\right] \\
&= \mathbb{E}_{\bar{J}_t,\bar{J}_{t+\tau}\sim\mathbb{C}_{N,K}}\left[\frac{1}{K^2}\sum_{i=1}^K\sum_{j=1}^K \mathbb{E}\left[Z_{t+\sum_{i'=1}^i d_{i'}} Z_{t+\tau+\sum_{j'=1}^j d_{j'}}\middle|\bar{J}_t, \bar{J}_{t+\tau}\right]\right] \\
&= \mathbb{E}_{\bar{J}_t,\bar{J}_{t+\tau}\sim\mathbb{C}_{N,K}}\left[\frac{1}{K^2}\sum_{i=1}^K\sum_{j=1}^K R_Z\left(\sum_{i'=1}^i d_{i'} - \sum_{j'=1}^j d_{j'} + \tau\right)\right].
\end{aligned}$$

Recalling Definition 1 of τ'_K we get

$$R_Y(\tau) = \mathbb{E}_{\tau'_K}\left[R_Z\left(\tau'_K\right)\right].$$

□

B.3. Proof of Lemma 4

Proof. Let $\bar{J}_t \subset \{1, \dots N\}$ and $\bar{J}_{t+\tau} \subset \{1, \dots N\}$ be subsets of K indices each. We saw in Section B.2 that we can move from these two subsets into the set of all possible differences $\tilde{J}_\tau$. Recall that we defined $\tau' = \tau + n - m$ to be the time difference between each couple of indices from $\bar{J}_t$ and $\bar{J}_{t+\tau}$. Note that $\tau' \in \tilde{J}_\tau$ where $\tilde{J}_\tau = \{\tau - N + 1, \dots, \tau + N - 1\}$.

Here we consider the "unordered sampling without replacement" (described in section 2.2) for sampling $\bar{J}_t$ and $\bar{J}_{t+\tau}$ and we would like to calculate the probability distribution for each time difference τ', that is $P(\tau')$. We have total of $K^2 \cdot \left(\binom{N}{K}\right)^2$

such differences since we have $\binom{N}{K}$ possible permutations for each batch and in each permutation we have K time elements. We define $d = \tau' - \tau$, therefore $-N+1 \leq d \leq N-1$. We now can calculate $P(\tau')$:

$$\begin{aligned}
P(\tau') &= \frac{\frac{K^2}{(N-K)^2} \cdot \frac{2}{1} \cdot \frac{3}{2} \cdots \frac{N-|\tau'-\tau|}{N-|\tau'-\tau|-1} \cdot \left(\binom{N-1}{K}\right)^2}{K^2 \cdot \left(\binom{N}{K}\right)^2} \\
&\overset{1}{=} \frac{\frac{K^2}{(N-K)^2} \cdot \frac{2}{1} \cdot \frac{3}{2} \cdots \frac{N-|d|}{N-|d|-1} \cdot \left(\binom{N-1}{K}\right)^2}{K^2 \cdot \left(\binom{N}{K}\right)^2} \\
&\overset{2}{=} \frac{(N-|d|) \cdot K! \cdot K! \cdot (N-K)! \cdot (N-K)! \cdot (N-1)! \cdot (N-1)!}{(N-K)^2 \cdot N! \cdot N! \cdot K! \cdot K! \cdot (N-1-K)! \cdot (N-1-K)!} \\
&\overset{3}{=} \frac{(N-|d|) \cdot (N-K)^2}{(N-K)^2 \cdot N^2} \\
&\overset{4}{=} \frac{N-|d|}{N^2},
\end{aligned}$$

where in (1) we substitute $\tau' - \tau = d$. In (2) we canceled similar elements in the denominator and numerator and we also wrote explicitly the binomial terms. In (3) and (4) we again canceled similar elements in the denominator and numerator. Notice that this probability formula is relevant only for $\tau - N + 1 \leq \tau' \leq \tau + N - 1$ and other values of τ' can not be reached form combining these two batches. Therefore, $P(\tau') = 0$ for $\tau - N + 1 > \tau'$ and $\tau' > \tau + N - 1$.

Interestingly, this proof show how parameter K is canceled out, meaning this time difference distribution is independent on K. In addition, we can observe that the resulting distribution can be considered as a convolution of two rectangles, which represents the time limits of each batch and the uniform sampling, and the resulting convolution, a triangle which represents $P(\tau')$. □

B.4. Proof of Corollary 1

Proof. Combining Theorem 1 Lemma 4 we get::

$$R_Y(\tau) = \mathbb{E}_{\tau'}\left[R_Z(\tau')\right] = \sum_{\tau'} P(\tau') R_Z(\tau') \overset{1}{=} \sum_{d=-N+1}^{N-1} \frac{N-|d|}{N^2} R_Z(d+\tau)$$

where in (1) we used $P(\tau')$ from Lemma 4 and changed the variables: $d = \tau' - \tau$ for $\tau - N + 1 \leq \tau' \leq \tau + N - 1$.

Similar development can be done to $C_Y(\tau)$. □

C. Proof of Theorem 2: Average reward and critic convergence

Proof. Recall that our TD-error update Algorithm 1 is defined as $\delta(O_j) = r(S_j, A_j) - \eta + \phi(S'_j)^\top w - \phi(S_j)^\top w$, where $O_j = \{S_j, A_j, r(S_j, A_j), S'_j\}$. In the critic update in Algorithm 1 we use an empirical mean of TD-errors of several sampled observations, denoted as $\{O_j\}_{j \in J}$. Then, the critic update is defined as

$$w' = w + \alpha^w \frac{1}{K} \sum_{j \in J} \delta(O_j) \phi(S_j).$$

where J is a random subset of K samples from RB with size N. Using the definition of the sampled random K indices $\bar{J}$, instead of J, we can write the update as:

$$w' = w + \alpha^w \frac{1}{K} \sum_{n \in \bar{J}} \delta(O_{t-n+1}) \phi(S_{t-n+1}).$$

In this proof we follow the proof of Lemma 5 in Bhatnagar et al. (2009). Observe that the average reward and critic updates

from Algorithm 1 can be written as

$$\eta_{t+1} = \eta_t + \alpha_t^\eta \left(F_t^t + M_{t+1}^\eta\right) \tag{C.1}$$

$$w_{t+1} = v_t + \alpha_t^w \left(F_t^w + M_{t+1}^w\right), \tag{C.2}$$

where

$$F_t^\eta \triangleq \mathbb{E}_{\bar{J}_t \sim \mathbb{C}_{N,K}, \{O_{t-n+1}\}_{n\in\bar{J}_t} \sim \text{RB}_t} \left[\frac{1}{K} \sum_{n\in\bar{J}_t} r(S_{t-n+1}, A_{t-n+1}) - \eta \middle| \mathcal{F}_t \right]$$

$$M_{t+1}^\eta \triangleq \left(\frac{1}{K} \sum_{n\in\bar{J}_t} r(S_{t-n+1}, A_{t-n+1}) - \eta_t \right) - F_t^\eta$$

$$F_t^w \triangleq \mathbb{E}_{\bar{J}_t \sim \mathbb{C}_{N,K}, \{O_{t-n+1}\}_{n\in\bar{J}_t} \sim \text{RB}_t} \left[\frac{1}{K} \sum_{n\in\bar{J}_t} \delta(O_{t-n+1})\phi(S_{t-n+1}) \middle| \mathcal{F}_t \right]$$

$$M_{t+1}^w \triangleq \frac{1}{K} \sum_{n\in\bar{J}_t} \delta(O_{t-n+1})\phi(S_{t-n+1}) - F_t^w$$

and $\mathcal{F}_t$ is a σ-algebra defined as $\mathcal{F}_t \triangleq \{\eta_\tau, w_\tau, M_\tau^\eta, M_\tau^w : \tau \leq t\}$.

We use Theorem 2.2 of Borkar & Meyn (2000) to prove convergence of these iterates. Briefly, this theorem states that given an iteration as in (C.1) and (C.2), these iterations are bounded w.p.1 if

Assumption 6. *1. F_t^η and F_t^w are Lipschitz, the functions $F_\infty(\eta) = \lim_{\sigma\to\infty} F^\eta(\sigma\eta)/\sigma$ and $F_\infty(w) = \lim_{\sigma\to\infty} F^w(\sigma w)/\sigma$ are Lipschitz, and $F_\infty(\eta)$ and $F_\infty(w)$ are asymptotically stable in the origin.*

2. The sequences M_{t+1}^η and M_{t+1}^w are martingale difference noises and for some C_0^η, C_0^w

$$\mathbb{E}\left[(M_{t+1}^\eta)^2 \middle| \mathcal{F}_t\right] \leq C_0^\eta (1 + \|\eta_t\|^2)$$

$$\mathbb{E}\left[(M_{t+1}^w)^2 \middle| \mathcal{F}_t\right] \leq C_0^w (1 + \|w_t\|^2).$$

We begin with the average reward update in (C.1). The ODE describing its asymptotic behavior corresponds to

$$\dot{\eta} = \mathbb{E}_{\bar{J}\sim\mathbb{C}_{N,K}, \{O_{t-n+1}\}_{n\in\bar{J}}\sim\text{RB}} \left[\frac{1}{K} \sum_{n\in\bar{J}_t} r(S_{t-n+1}, A_{t-n+1}) - \eta \right] \triangleq F^\eta. \tag{C.3}$$

F^η is Lipschitz continuous in η. The function $F_\infty(\eta)$ exists and satisfies $F_\infty(\eta) = -\eta$. The origin is an asymptotically stable equilibrium for the ODE $\dot{\eta} = F_\infty(\eta)$ and the related Lyapunov function is given by $\eta^2/2$.

For the critic update, consider the ODE

$$\dot{w} = \mathbb{E}_{\bar{J}\sim\mathbb{C}_{N,K}, \{O_{t-n+1}\}_{n\in\bar{J}}\sim\text{RB}} \left[\frac{1}{K} \sum_{n\in\bar{J}} \delta(O_{t-n+1})\phi(S_{t-n+1}) \right] \triangleq F^w$$

In Lemma 5 we show that this ODE can be written as

$$\dot{w} = \Phi^\top C_\theta \Phi w + \Phi^\top b_\theta, \tag{C.4}$$

where C_θ and b_θ are defined in (7). F^w is Lipschitz continuous in w and $F_\infty(w)$ exists and satisfies $F_\infty(w) = \Phi^\top C_\theta \Phi w$. Consider the system

$$\dot{w} = F_\infty(w) \tag{C.5}$$

In assumption 4 we assume that $\Phi w \neq e$ for every $w \in \mathbb{R}^d$. Therefore, the only asymptotically stable equilibrium for (C.5) is the origin (see the explanation in the proof of Lemma 5 in Bhatnagar et al. (2009)). Therefore, for all $t \geq 0$

$$\mathbb{E}\left[(M_{t+1}^{\eta})^2 \middle| \mathcal{F}_t\right] \leq C_0^{\eta}(1 + \|\eta_t\|^2 + \|w_t\|^2)$$

$$\mathbb{E}\left[(M_{t+1}^{w})^2 \middle| \mathcal{F}_t\right] \leq C_0^{w}(1 + \|\eta_t\|^2 + \|w_t\|^2)$$

for some $C_0^{\eta}, C_0^{w} < \infty$. M_t^{η} can be directly seen to be uniformly bounded almost surely. Thus, Assumptions (A1) and (A2) of Borkar & Meyn (2000) are satisfied for the average reward, TD-error, and critic updates. From Theorem 2.1 of Borkar & Meyn (2000), the average reward, TD-error, and critic iterates are uniformly bounded with probability one. Note that when $t \to \infty$, (C.3) has η_θ defined as in (2) as its unique globally asymptotically stable equilibrium with $V_2(\eta) = (\eta - \eta_\theta)^2$ serving as the associated Lyapunov function.

Next, suppose that $w = w^\pi$ is a solution to the system $\Phi^\top C_\theta \Phi w = 0$. Under Assumption 4, using the same arguments as in the proof of Lemma 5 in Bhatnagar et al. (2009), w^π is the unique globally asymptotically stable equilibrium of the ODE (C.4). Assumption 6 is now verified and under Assumption 5, the claim follows from Theorem 2.2, pp. 450 of (Borkar & Meyn, 2000). □

C.1. Proof of Lemma 5

Proof. We compute the expectation of the critic update with linear function approximation according to Algorithm 1. In this proof, we focus on the "Unordered sampling without replacement" strategy for sampling batch of K transitions from the replay buffer (see Section 2.2 for this strategy probability distribution). Recall that n is a position in the RB and it corresponds to transition $O_{t-n+1} = (S_{t-n+1}, A_{t-n+1}, S'_{t-n+1})$. We will use the notation of $\bar{J} \subset \{1, \ldots, n, \ldots, N\}$ to refer the K indices sampled batches. In addition we will use the following observations:

$$P(n|\bar{J}, n \in \bar{J}) = \frac{1}{K}, \quad P(n|\bar{J}, n \notin \bar{J}) = 0 \tag{C.6}$$

$$P(n \in \bar{J}) = \frac{K}{N}, \quad P(n \notin \bar{J}) = 1 - \frac{K}{N}$$

$$P(n|\bar{J}) = P(n \in \bar{J}) \cdot P(n|\bar{J}, n \in \bar{J}) + P(n \notin \bar{J}) \cdot P(n|\bar{J}, n \notin \bar{J}) = \frac{K}{N}\frac{1}{K} + 0 = \frac{1}{N} \tag{C.7}$$

$$P(\bar{J}) = \frac{1}{\binom{N}{K}} \tag{C.8}$$

$$|\mathbb{C}_{N,K}| = \binom{N}{K} \tag{C.9}$$

Now we can compute the desired expectation:

$$
\begin{aligned}
&\mathbb{E}_{\bar{J}_t\sim\mathbb{C}_{N,K},\{O_{t-n+1}\}_{n\in\bar{J}_t}\sim\mathrm{RB}_t}\left[\frac{1}{K}\sum_{n\in\bar{J}_t}\delta(O_{t-n+1})\phi(S_{t-n+1})\right] \\
&= \mathbb{E}_{\bar{J}_t\sim\mathbb{C}_{N,K}}\left[\mathbb{E}_{\{O_{t-n+1}\}_{n\in\bar{J}_t}\sim\mathrm{RB}_t}\left[\frac{1}{K}\sum_{n\in\bar{J}_t}\delta(O_{t-n+1})\phi(S_{t-n+1})\middle|\bar{J}_t\right]\right] \\
&\overset{(C.6)}{=} \mathbb{E}_{\bar{J}_t\sim\mathbb{C}_{N,K}}\left[\mathbb{E}_{\{O_{t-n+1}\}_{n\in\bar{J}_t}\sim\mathrm{RB}_t}\left[\mathbb{E}_{n\sim\bar{J}_t}\left[\delta(O_{t-n+1})\phi(S_{t-n+1})\right]\middle|\bar{J}_t\right]\right] \\
&\overset{1}{=} \mathbb{E}_{\bar{J}_t\sim\mathbb{C}_{N,K}}\left[\mathbb{E}_{n\sim\bar{J}_t}\left[\mathbb{E}_{O_{t-n+1}}\left[\delta(O_{t-n+1})\phi(S_{t-n+1})\right]\right]\middle|\bar{J}_t\right] \\
&\overset{2}{=} \sum_{\bar{J}_t\in\mathbb{C}_{N,K}}P(\bar{J}_t)\sum_{n=1}^{N}P(n|\bar{J}_t)\mathbb{E}_{O_{t-n+1}}\left[\delta(O_{t-n+1})\phi(S_{t-n+1})\right] \\
&\overset{(C.7),(C.8)}{=} \sum_{\bar{J}_t\in\mathbb{C}_{N,K}}\frac{1}{\binom{N}{K}}\sum_{n=1}^{N}\frac{1}{N}\mathbb{E}_{O_{t-n+1}}\left[\delta(O_{t-n+1})\phi(S_{t-n+1})\right] \\
&\overset{(C.9)}{=} \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{O_{t-n+1}}\left[\delta(O_{t-n+1})\phi(S_{t-n+1})\right] \\
&\overset{3}{=} \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{S_{t-n+1},A_{t-n+1},S'_{t-n+1}}\left[\left(r(S_{t-n+1},A_{t-n+1})-\eta+\phi(S'_{t-n+1})^\top w-\phi(S_{t-n+1})^\top w\right)\phi(S_{t-n+1})\right]
\end{aligned}
\tag{C.10}
$$

where in (1) we used the linearity of expectation and that given the sampled batch $\bar{J}$, the transitions tuples $\{O_{t-n_k+1}\}_{k=1}^{K}$ are sampled independently. In (2) we wrote expectations explicitly and in (3) we used the definition of the TD-error in (4).

Next, for time $t-n+1$ where $1\le n\le N$, we define the induced MC with a corresponding policy parameter θ_{t-n+1}. For this parameter, we denote the corresponding state distribution vector ρ_{t-n+1} and a transition matrix P_{t-n+1} (both induced by the policy $\pi_{\theta_{t-n+1}}$. In addition, we define the following diagonal matrix $D_{t-n+1}\triangleq\mathrm{diag}(\rho_{t-n+1})$. Similarly to (Bertsekas & Tsitsiklis, 1996) Lemma 6.5, pp.298, we can substitute the inner expectation

$$
\begin{aligned}
&\mathbb{E}_{S_{t-n+1},A_{t-n+1},S'_{t-n+1}}\left[\left(r(S_{t-n+1},A_{t-n+1})-\eta+\phi(S'_{t-n+1})^\top w-\phi(S_{t-n+1})^\top w\right)\phi(S_{t-n+1})\right] \\
&=\Phi^\top D_{t-n+1}\left(P_{t-n+1}-I\right)\Phi w+\Phi^\top D_{t-n+1}(r_{t-n+1}-\eta_\theta e),
\end{aligned}
\tag{C.11}
$$

where I is the $|\mathcal{S}|\times|\mathcal{S}|$ identity matrix, e in $|\mathcal{S}|\times 1$ vector of ones and r_{t-n+1} is a $|\mathcal{S}|\times 1$ vector defined as $r_{t-n+1}(s)=\sum_a\pi_{\theta_{t-n+1}}(A|S)r(S,A)$. Combining equations (6), (C.10) and (C.11) yields

$$
\frac{1}{N}\sum_{n=1}^{N}\left(\Phi^\top D_{t-n+1}\left(P_{t-n+1}-I\right)\Phi w+\Phi^\top D_{t-n+1}(r_{t-n+1}-\eta_\theta e)\right)=\Phi^\top C_t\Phi w+\Phi^\top b_t,
\tag{C.12}
$$

In the limit, $t\to\infty$ and $\rho_{t-n+1}\to\mu_\theta$ for all index n. Using C_θ and b_θ defined in (7), (C.10) can be expressed as

$$
\mathbb{E}_{\bar{J}_t\sim\mathbb{C}_{N,K},\{O_{t-n+1}\}_{n\in\bar{J}_t}\sim\mathrm{RB}_t}\left[\frac{1}{K}\sum_{n\in\bar{J}_t}\delta(O_{t-n+1})\phi(S_{t-n+1})\right]=\Phi^\top C_\theta\Phi w+\Phi^\top b_\theta.
\tag{C.13}
$$

□

D. Proof of Theorem 3: Actor convergence

Proof. Recall that our TD-error update in Algorithm 1 is defined as $\delta(O_j) = r(S_j, A_j) - \eta + \phi(S'_j)^\top w - \phi(S_j)^\top w$, where $O_j = \{S_j, A_j, r(S_j, A_j), S'_j\}$. In the actor update in Algorithm 1 we use an empirical mean of TD-errors of several sampled observations, denoted as $\{O_j\}_{j\in J}$. Then, the actor update is defined as

$$\theta' = \Gamma\left(\theta - \alpha^\theta \frac{1}{K}\sum_{j\in J}\delta(O_j)\nabla\log\pi_\theta(A_j|S_j)\right).$$

where J is a random subset of K samples from RB with size N. Using the definition of the sampled random K indices $\bar{J}$, instead of J, we can write the update as:

$$\theta' = \Gamma\left(\theta - \alpha^\theta \frac{1}{K}\sum_{n\in\bar{J}}\delta(O_{t-n+1})\nabla\log\pi_\theta(A_{t-n+1}|S_{t-n+1})\right).$$

In this proof we follow the proof of Theorem 2 in Bhatnagar et al. (2009). Let $O = \{S, A, S'\}$ and let $\delta^\pi(O) = r(S, A) - \eta + \phi(S')^\top w^\pi - \phi(S)^\top w^\pi$, where w^π is the convergent parameter of the critic recursion with probability one (see its definition in the proof for Theorem 2). Observe that the actor parameter update from Algorithm 1 can be written as

$$\begin{aligned}\theta_{t+1} &= \Gamma\Big(\theta_t - \alpha_t^\theta\big(\delta(O)\nabla_\theta\log\pi_\theta(A|S) + F_t^\theta - F_t^\theta + N_t^{\theta_t} - N_t^{\theta_t}\big)\Big)\\ &= \Gamma\Big(\theta_t - \alpha_t^\theta\big(M_{t+1}^\theta + (F_t^\theta - N_t^{\theta_t}) + N_t^{\theta_t}\big)\Big)\end{aligned}$$

where

$$\begin{aligned}F_t^\theta &\triangleq \mathbb{E}_{\bar{J}_t\sim\mathbb{C}_{N,K},\{O_{t-n+1}\}_{n\in\bar{J}_t}\sim\mathrm{RB}_t}\left[\frac{1}{K}\sum_{n\in\bar{J}_t}\delta(O_{t-n+1})\nabla_\theta\log\pi_\theta(A_{t-n+1}|S_{t-n+1})\middle|\mathcal{F}_t\right]\\ M_{t+1}^\theta &\triangleq \frac{1}{K}\sum_{n\in\bar{J}_t}\delta(O_{t-n+1})\nabla_\theta\log\pi_\theta(A_{t-n+1}|S_{t-n+1}) - F_t^\theta\\ N_t^\theta &\triangleq \mathbb{E}_{\bar{J}_t\sim\mathbb{C}_{N,K},\{O_{t-n+1}\}_{n\in\bar{J}_t}\sim\mathrm{RB}_t}\left[\frac{1}{K}\sum_{n\in\bar{J}_t}\delta^{\pi_\theta}(O_{t-n+1})\nabla_\theta\log\pi_\theta(A_{t-n+1}|S_{t-n+1})\middle|\mathcal{F}_t\right]\end{aligned}$$

and $\mathcal{F}_t$ is a σ-algebra defined as $\mathcal{F}_t \triangleq \{\eta_\tau, w_\tau, \theta_\tau, M_\tau^\eta, M_\tau^w, M_\tau^\theta : \tau \le t\}$.

Since the critic converges along the faster timescale, from Theorem 2 it follows that $F_t^\theta - N_t^{\theta_t} = o(1)$. Now, let

$$M_2(t) = \sum_{r=0}^{t-1}\alpha_r^\theta M_{r+1}^\theta, t \ge 1.$$

The quantities $\delta(O)$ can be seen to be uniformly bounded since from the proof in Theorem 2, $\{\eta_t\}$ and $\{w_t\}$ are bounded sequences. Therefore, using Assumption 5, $\{M_2(t)\}$ is a convergent martingale sequence (Bhatnagar & Kumar, 2004).

Consider the actor update along the slower timescale corresponding to α_t^θ in Algorithm 1. Let $w(\cdot)$ be a vector field on a set Θ. Define another vector field: $\hat{\Gamma}\big(w(y)\big) = \lim_{0<\eta\to 0}\left(\frac{\Gamma\left(y+\eta w(y)\right)-y}{\eta}\right)$. In case this limit is not unique, we let $\hat{\Gamma}\big(w(y)\big)$ be the set of all possible limit points (see pp. 191 of (Kushner & Clark, 2012)). Consider now the ODE

$$\dot{\theta} = \hat{\Gamma}\left(-\mathbb{E}_{\bar{J}_t\sim\mathbb{C}_{N,K},\{O_{t-n+1}\}_{n\in\bar{J}_t}\sim\mathrm{RB}_t}\left[\frac{1}{K}\sum_{n\in\bar{J}_t}\delta^{\pi_\theta}(O_{t-n+1})\nabla_\theta\log\pi_\theta(A_{t-n+1}|S_{t-n+1})\right]\right) \tag{D.1}$$

Substituting the result from Lemma 6, the above ODE is analogous to

$$\dot{\theta} = \hat{\Gamma}(-\nabla_\theta \eta_\theta + \xi^{\pi_\theta}) = \hat{\Gamma}\big(-N_t^\theta\big) \tag{D.2}$$

where $\xi^{\pi_\theta} = \sum_S \mu_\theta(S)\Big(\phi(S)^\top \nabla_\theta w^{\pi_\theta} - \nabla_\theta \bar{V}^{\pi_\theta}(S)\Big)$. Consider also an associated ODE:

$$\dot{\theta} = \hat{\Gamma}\big(-\nabla_\theta \eta_\theta\big) \tag{D.3}$$

We now show that $h_1(\theta_t) \triangleq -N_t^{\theta_t}$ is Lipschitz continuous. Here $w^{\pi_{\theta_t}}$ corresponds to the weight vector to which the critic update converges along the faster timescale when the corresponding policy is π_{θ_t} (see Theorem 2). Note that $\mu_\theta(S), S \in \mathcal{S}$ is continuously differentiable in θ and have bounded derivatives. Also, $\bar{\eta}_{\theta_t}$ is continuously differentiable as well and has bounded derivative as can also be seen from (2). Further, $w^{\pi_{\theta_t}}$ can be seen to be continuously differentiable with bounded derivatives. Finally, $\nabla^2 \pi_{\theta_t}(A|S)$ exists and is bounded. Thus $h_1(\theta_t)$ is a Lipschitz continuous function and the ODE (D.1) is well posed.

Let $\mathcal{Z}$ denote the set of asymptotically stable equilibria of (D.3) i.e., the local minima of η_θ, and let $\mathcal{Z}^\epsilon$ be the ϵ-neighborhood of $\mathcal{Z}$. To complete the proof, we are left to show that as $\sup_\theta \|\xi^{\pi_\theta}\| \to 0$ (viz. $\delta \to 0$), the trajectories of (D.2) converge to those of (D.3) uniformly on compacts for the same initial condition in both. This claim follows the same arguments as in the proof of Theorem 2 in Bhatnagar et al. (2009). □

D.1. Proof of Lemma 6

Proof. We compute the required expectation with linear function approximation according to Algorithm 1. Following the same steps when proving the expectation for the critic in Section C.1, we have:

$$\begin{aligned}
&\mathbb{E}_{\bar{J}_t \sim \mathbb{C}_{N,K}, \{O_{t-n+1}\}_{n\in\bar{J}_t} \sim \mathrm{RB}_t}\left[\frac{1}{K}\sum_{n\in\bar{J}_t} \delta^{\pi_\theta}(O_{t-n+1}) \nabla_\theta \log \pi_\theta(A_{t-n+1}|S_{t-n+1})\right] \\
&= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{S_{t-n+1}, A_{t-n+1}, S'_{t-n+1}}\left[\left(r(S_{t-n+1}, A_{t-n+1}) - \eta + \phi(S'_{t-n+1})^\top w - \phi(S_{t-n+1})^\top w\right) \nabla_\theta \log \pi_\theta(A_{t-n+1}|S_{t-n+1})\right]
\end{aligned}$$

Recall the definition of the state distribution vector ρ_{t-n+1} in Section 4.4. In the limit, $t \to \infty$ and $\rho_{t-n+1} \to \mu_\theta$ for all index n, then:

$$\begin{aligned}
&\mathbb{E}_{\bar{J}_t \sim \mathbb{C}_{N,K}, \{O_{t-n+1}\}_{n\in\bar{J}_t} \sim \mathrm{RB}_t}\left[\frac{1}{K}\sum_{n\in\bar{J}_t} \delta^{\pi_\theta}(O_{t-n+1}) \nabla_\theta \log \pi_\theta(A_{t-n+1}|S_{t-n+1})\right] \\
&= \sum_{S\in\mathcal{S}} \mu_\theta(S) \sum_{A\in\mathcal{A}} \pi_\theta(A|S)\left(r(S,A) - \eta_\theta + \sum_{S'\in\mathcal{S}} P(S'|S,A)\phi(S')^\top w^{\pi_\theta} - \phi(S)^\top w^{\pi_\theta}\right) \nabla_\theta \log \pi_\theta(A|S)
\end{aligned}$$

We define now the following term:

$$\bar{V}^{\pi_\theta}(S) = \sum_{A\in\mathcal{A}} \pi_\theta(A|S) \bar{Q}^{\pi_\theta}(S,A) = \sum_{A\in\mathcal{A}} \pi_\theta(A|S)\left(r(S,A) - \eta_\theta + \sum_{S'\in\mathcal{S}} P(S'|S,A)\phi(S')^\top w^{\pi_\theta}\right), \tag{D.4}$$

where $\bar{V}^{\pi_\theta}(S)$ and $\bar{Q}^{\pi_\theta}(S,A)$ correspond to policy π_θ. Note that here, the convergent critic parameter w^{π_θ} is used. Let's look at the gradient of (D.4):

$$\begin{aligned}
\nabla_\theta \bar{V}^{\pi_\theta}(S) &= \nabla_\theta \left(\sum_{A \in \mathcal{A}} \pi_\theta(A|S) \bar{Q}^{\pi_\theta}(S, A) \right) \\
&= \sum_{A \in \mathcal{A}} \nabla_\theta \pi_\theta(A|S) \left(r(S, A) - \eta_\theta + \sum_{S' \in \mathcal{S}} P(S'|S, A) \phi(S')^\top w^{\pi_\theta} \right) \\
&+ \sum_{A \in \mathcal{A}} \pi_\theta(A|S) \left(-\nabla_\theta \eta_\theta + \sum_{S' \in \mathcal{S}} P(S'|S, A) \phi(S')^\top \nabla_\theta w^{\pi_\theta} \right) \\
&= \sum_{A \in \mathcal{A}} \nabla_\theta \pi_\theta(A|S) \left(r(S, A) - \eta_\theta + \sum_{S' \in \mathcal{S}} P(S'|S, A) \phi(S')^\top w^{\pi_\theta} \right) \\
&- \nabla_\theta \eta_\theta + \sum_{A \in \mathcal{A}} \pi_\theta(A|S) \sum_{S' \in \mathcal{S}} P(S'|S, A) \phi(S')^\top \nabla_\theta w^{\pi_\theta}
\end{aligned}$$

Summing both sides over the stationary distribution μ_θ

$$\begin{aligned}
\sum_S \mu_\theta(S) \nabla_\theta \bar{V}^{\pi_\theta}(S) &= \sum_S \mu_\theta(S) \sum_{A \in \mathcal{A}} \nabla_\theta \pi_\theta(A|S) \left(r(S, A) - \eta_\theta + \sum_{S' \in \mathcal{S}} P(S'|S, A) \phi(S')^\top w^{\pi_\theta} \right) \\
&+ \sum_S \mu_\theta(S) \left(-\nabla_\theta \eta_\theta + \sum_{A \in \mathcal{A}} \pi_\theta(A|S) \sum_{S' \in \mathcal{S}} P(S'|S, A) \phi(S')^\top \nabla_\theta w^{\pi_\theta} \right) \\
&= \mathbb{E}_{\bar{J}_t \sim \mathbb{C}_{N,K}, \{O_{t-n+1}\}_{n \in \bar{J}_t} \sim \mathrm{RB}_t} \left[\frac{1}{K} \sum_{n \in \bar{J}_t} \delta^{\pi_\theta}(O_{t-n+1}) \nabla_\theta \log \pi_\theta(A_{t-n+1}|S_{t-n+1}) \right] \\
&- \nabla_\theta \eta_\theta + \sum_S \mu_\theta(S) \sum_{A \in \mathcal{A}} \pi_\theta(A|S) \sum_{S' \in \mathcal{S}} P(S'|S, A) \phi(S')^\top \nabla_\theta w^{\pi_\theta}
\end{aligned}$$

Then:

$$\begin{aligned}
\nabla_\theta \eta_\theta &= \mathbb{E}_{\bar{J}_t \sim \mathbb{C}_{N,K}, \{O_{t-n+1}\}_{n \in \bar{J}_t} \sim \mathrm{RB}_t} \left[\frac{1}{K} \sum_{n \in \bar{J}_t} \delta^{\pi_\theta}(O_{t-n+1}) \nabla_\theta \log \pi_\theta(A_{t-n+1}|S_{t-n+1}) \right] \\
&+ \sum_S \mu_\theta(S) \left(\sum_{A \in \mathcal{A}} \pi_\theta(A|S) \sum_{S' \in \mathcal{S}} P(S'|S, A) \phi(S')^\top \nabla_\theta v^{\pi_\theta} - \nabla_\theta \bar{V}^{\pi_\theta}(S) \right).
\end{aligned}$$

Since μ_θ is a stationary distribution,

$$\begin{aligned}
\sum_S \mu_\theta(S) \sum_{A \in \mathcal{A}} \pi_\theta(A|S) \sum_{S' \in \mathcal{S}} P(S'|S, A) \phi(S')^\top \nabla_\theta w^{\pi_\theta} &= \sum_S \mu_\theta(S) \sum_{S' \in \mathcal{S}} P_\theta(S'|S) \phi(S')^\top \nabla_\theta w^{\pi_\theta} \\
&= \sum_{S'} \sum_S \mu_\theta(S) P_\theta(S'|S) \phi(S')^\top \nabla_\theta w^{\pi_\theta} \\
&= \sum_{S'} \mu_\theta(S') \phi(S')^\top \nabla_\theta w^{\pi_\theta},
\end{aligned}$$

Then,

$$\begin{aligned}
\nabla_\theta \eta_\theta &= \mathbb{E}_{\bar{J}_t \sim \mathbb{C}_{N,K}, \{O_{t-n+1}\}_{n \in \bar{J}_t} \sim \mathrm{RB}_t} \left[\frac{1}{K} \sum_{n \in \bar{J}_t} \delta^{\pi_\theta}(O_{t-n+1}) \nabla_\theta \log \pi_\theta(A_{t-n+1}|S_{t-n+1}) \right] \\
&+ \sum_S \mu_\theta(S) \Big(\phi(S)^\top \nabla_\theta w^{\pi_\theta} - \nabla_\theta \bar{V}^{\pi_\theta}(S) \Big)
\end{aligned}$$

The result follows immediately. □