

Analysis of Stochastic Processes through Replay Buffers

Shirli Di-Castro Shashua (Technion)

Shie Mannor (Technion, NVIDIA)

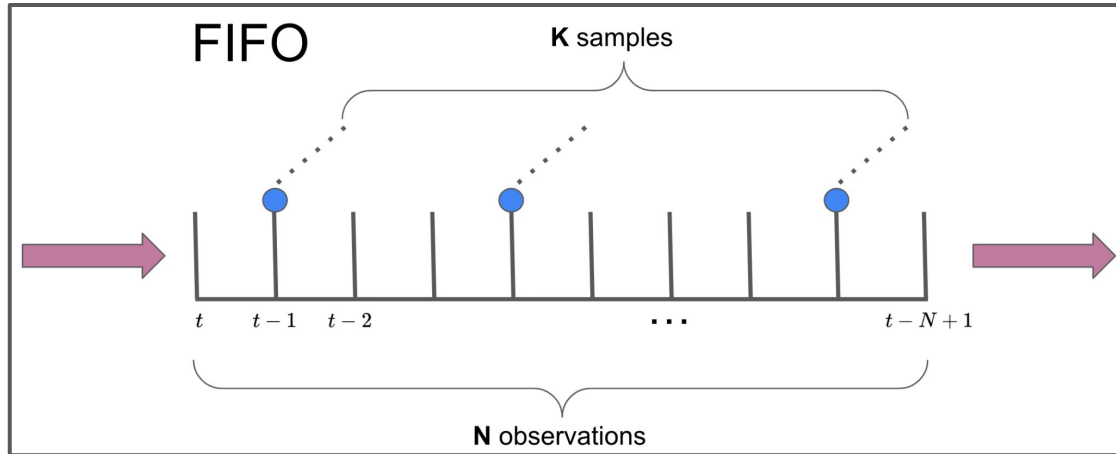
Dotan Di-Castro (Bosch center of AI)

ICML 2022



Replay Buffer (RB) Mechanism

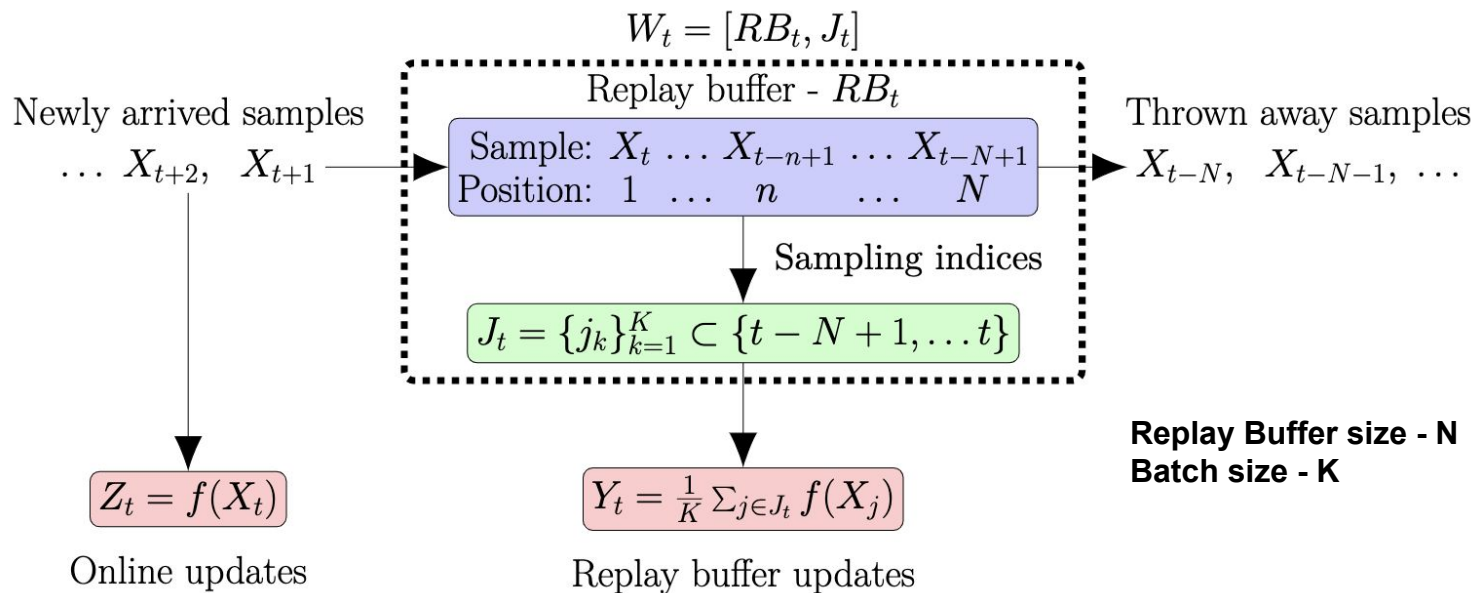
- First-in-First-out (FIFO) queue
- Storing last N transitions observed by an agent
- Sampling K transitions from the RB to calculate a gradient update



Our Contributions

- We define the stochastic processes that are involved in the Replay buffer mechanism
- We prove interesting properties for the RB processes: Markovity, stationarity and ergodicity
- We prove that the RB acts as a decorrelator
- We prove, for the first time, the asymptotic convergence of an RB-based actor critic algorithm

Involving stochastic processes in a replay buffer scheme:



Replay buffer properties:

Lemma 1 - **Stationarity:** X_t, J_t stationary $\rightarrow RB_t, Y_t$ stationary.

Lemma 2 - **Markovity:** X_t Markovian $\rightarrow RB_t, W_t$ Markovian.

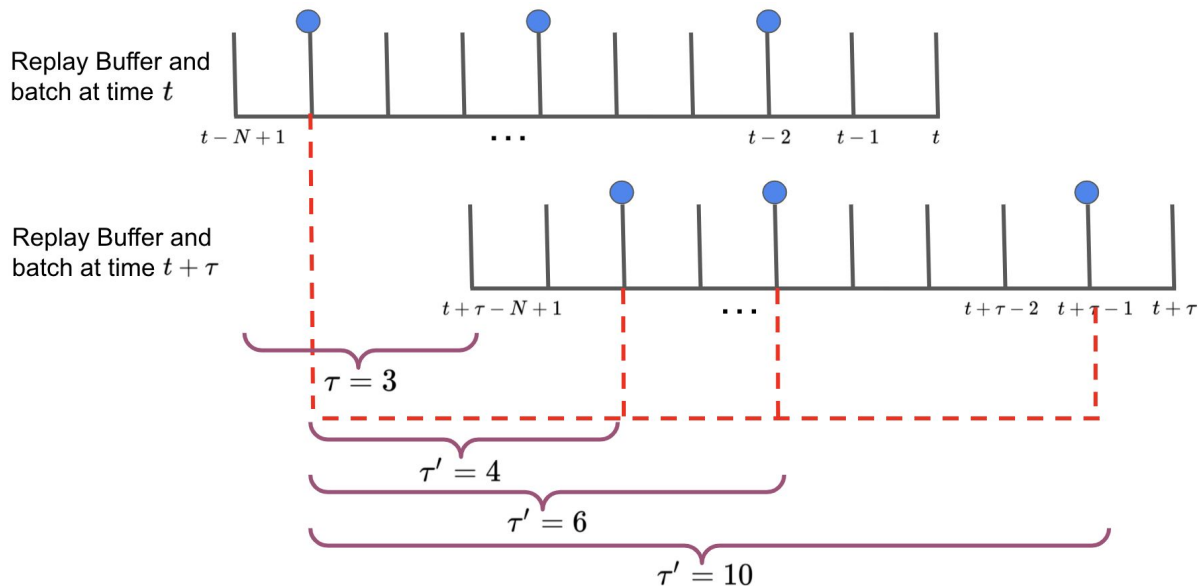
Lemma 3 - **Ergodicity:** X_t Markovian + ergodic $\rightarrow RB_t, W_t$ Markovian + ergodic.

Replay Buffer as a de-correlator

Auto-correlation definition: $R_X(\tau) = \mathbb{E}[X_t X_{t+\tau}]$

Theorem 1: The autocorrelation of Y_t process is an expectation over the autocorrelation of Z_t process.

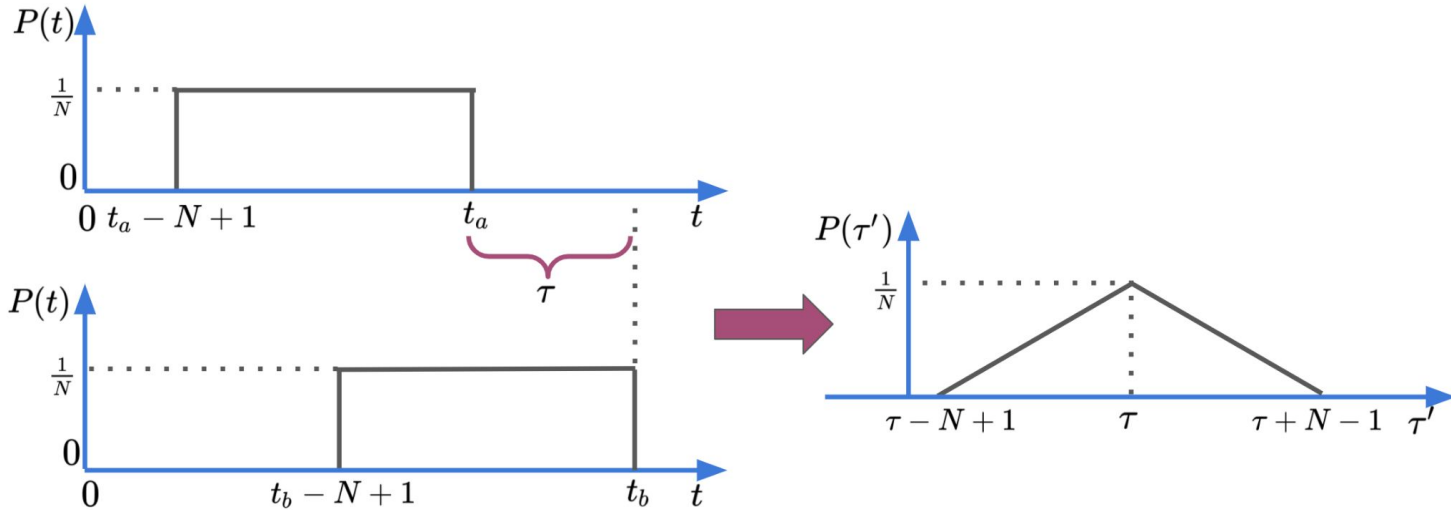
$$R_Y(\tau) = \mathbb{E}_{\tau'} [R_Z(\tau')]$$



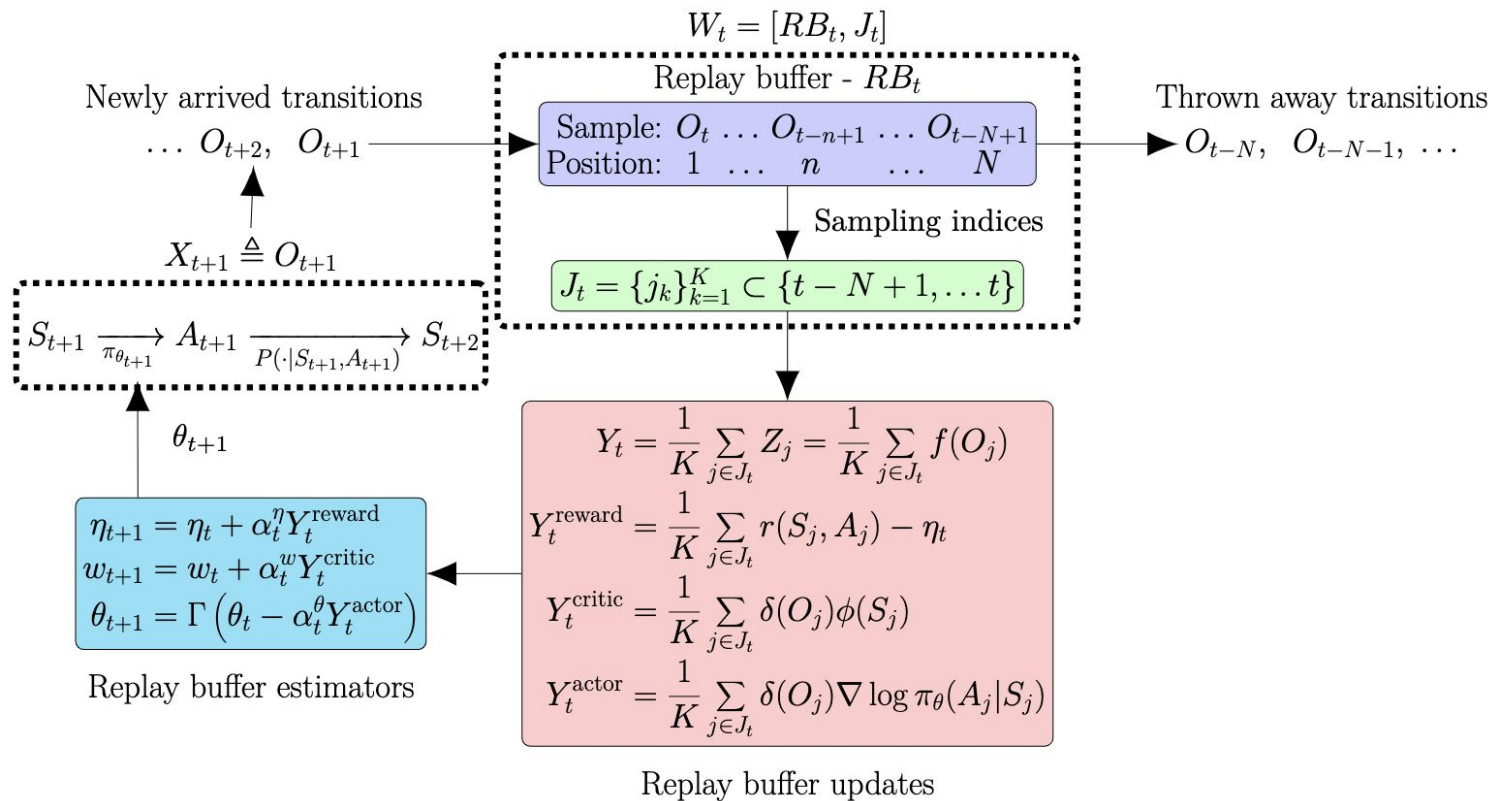
Replay Buffer as a de-correlator

Corollary 1: For batches sampled uniformly without replacement:

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



Replay Buffers in Reinforcement Learning



Convergence Proof for Linear Actor Critic with Replay Buffer Samples

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** $t = 0, \dots$ **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 from 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

$$\eta_{t+1} = \eta_t + \alpha_t^\eta \left(\frac{1}{K} \sum_{j \in J_t} r(S_j, A_j) - \eta_t \right)$$
 - 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 $\theta_{t+1} = \Gamma(\theta_t - \alpha_t^\theta \frac{1}{K} \sum_{j \in J_t} \delta(O_j) \nabla_\theta \log \pi_\theta(A_j | S_j))$
 - 11: **end for**
-

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 $\eta_t \rightarrow \eta_\theta$ and $w_t \rightarrow w^\pi$ with probability 1, where w^π is obtained as a unique solution to $\Phi^\top C_\theta \Phi w + \Phi^\top b_\theta = 0$.

Theorem 3. (Convergence of the actor)

Under Assumptions 1-5, given $\epsilon > 0, \exists \delta > 0$ such that for $\theta_t, t \geq 0$ obtained using Algorithm 1, if $\sup_{\theta_t} \|\xi^{\pi_{\theta_t}}\| < \delta$, then $\theta_t \rightarrow \mathcal{Z}^\epsilon$ as $t \rightarrow \infty$ with probability one.