

SPEQ: OFFLINE STABILIZATION PHASES FOR EFFICIENT Q-LEARNING IN HIGH UPDATE-TO-DATA RATIO REINFORCEMENT LEARNING



Romeo C.*¹, Macaluso G.*¹, Sestini A.², Bagdanov A. D.¹

¹Media Integration and Communication Center, University of Florence - ²SEED, Electronic Arts

¹{carlo.romeo, girolamo.macaluso, Andrew.bagdanov}@unifi.it, ²asestini@ea.com



Abstract

High update-to-data (UTD) ratio algorithms improve sample efficiency but are computationally expensive. We introduce **SPEQ**, a RL method that combines low-UTD online training with **periodic offline stabilization phases**, where Q-functions are fine-tuned using a fixed replay buffer. This reduces redundant updates on poor data and balance between sample and compute efficiency. SPEQ achieves **40–99% fewer gradient updates** and **27–78% less training time** than SOTA methods while matching or exceeding their performance.

Introduction

- Context:** **Sample efficiency** is a critical challenge in Reinforcement Learning (RL). Recent methods using a **high UTD ratio** excel at reusing collected experience, leading to improved performance.
- Problem:** High UTD methods dramatically **increase training times** and requires **heavier computational** demands.
- Goal:** **Preserve the sample efficiency** of high-UTD strategies while **reducing the computational overhead**.

SPEQ Method



Two-Phase Training.

Alternate between two training phases to balance efficiency and computational cost, like in **offline RL**



Online Phase — Low UTD

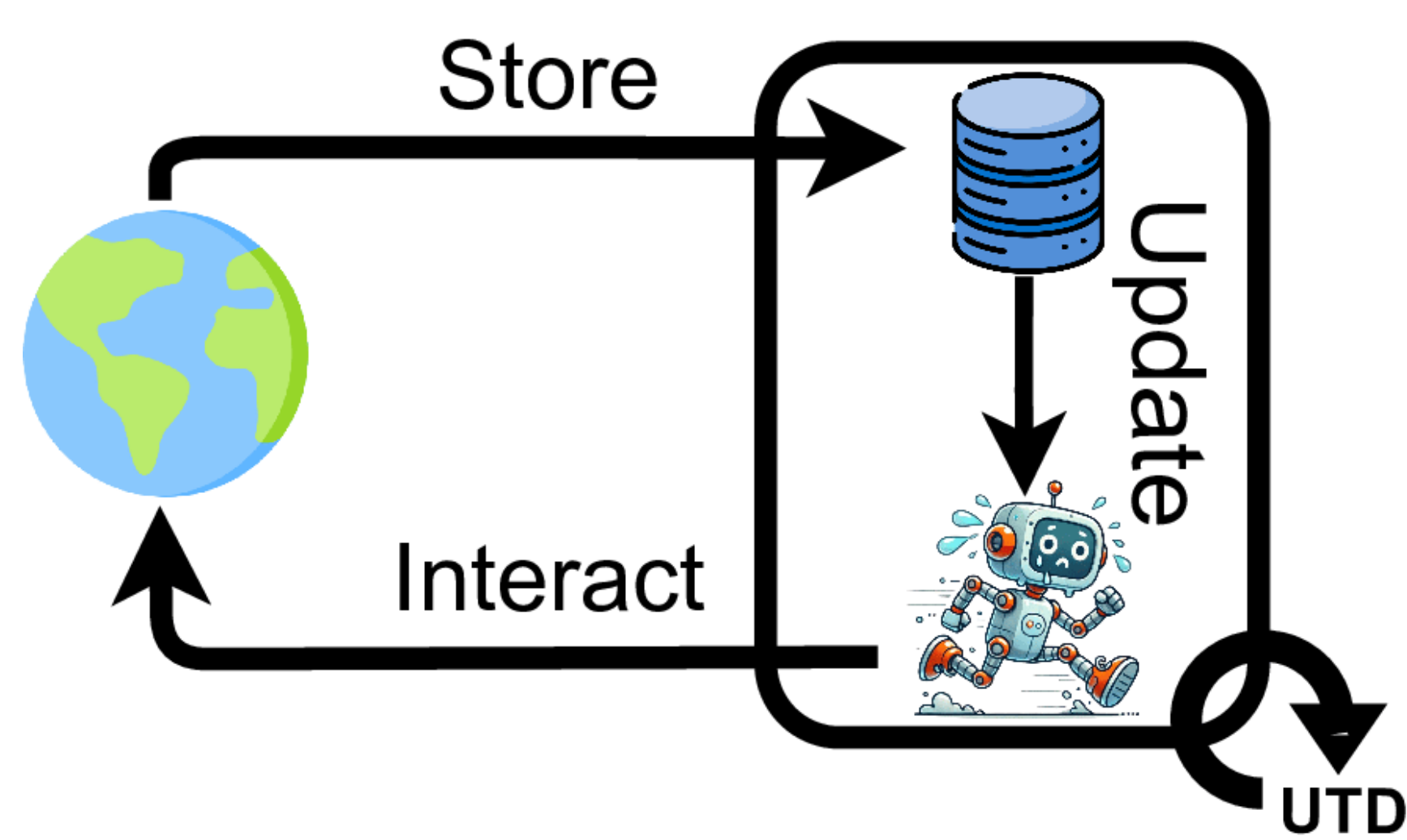
Interact with the environment in real-time and perform **lightweight updates** after each step. This keeps training responsive and stable.



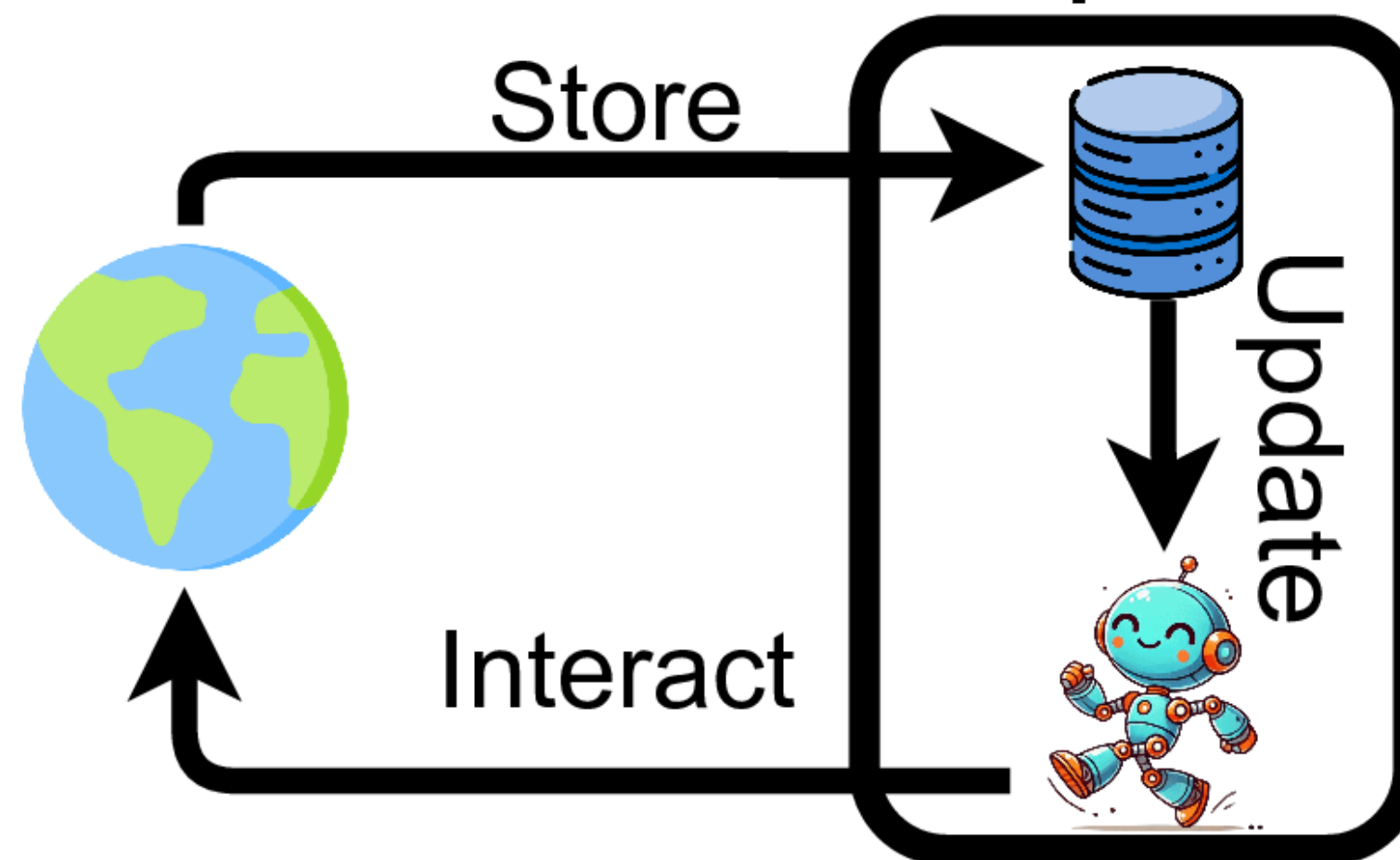
Offline Stabilization Phase — High UTD

Pause environment interaction and perform **intensive training** using only the replay buffer. This phase enables **high UTD updates** without additional data collection to stabilize the Q-Functions.

High UTD ratio approaches

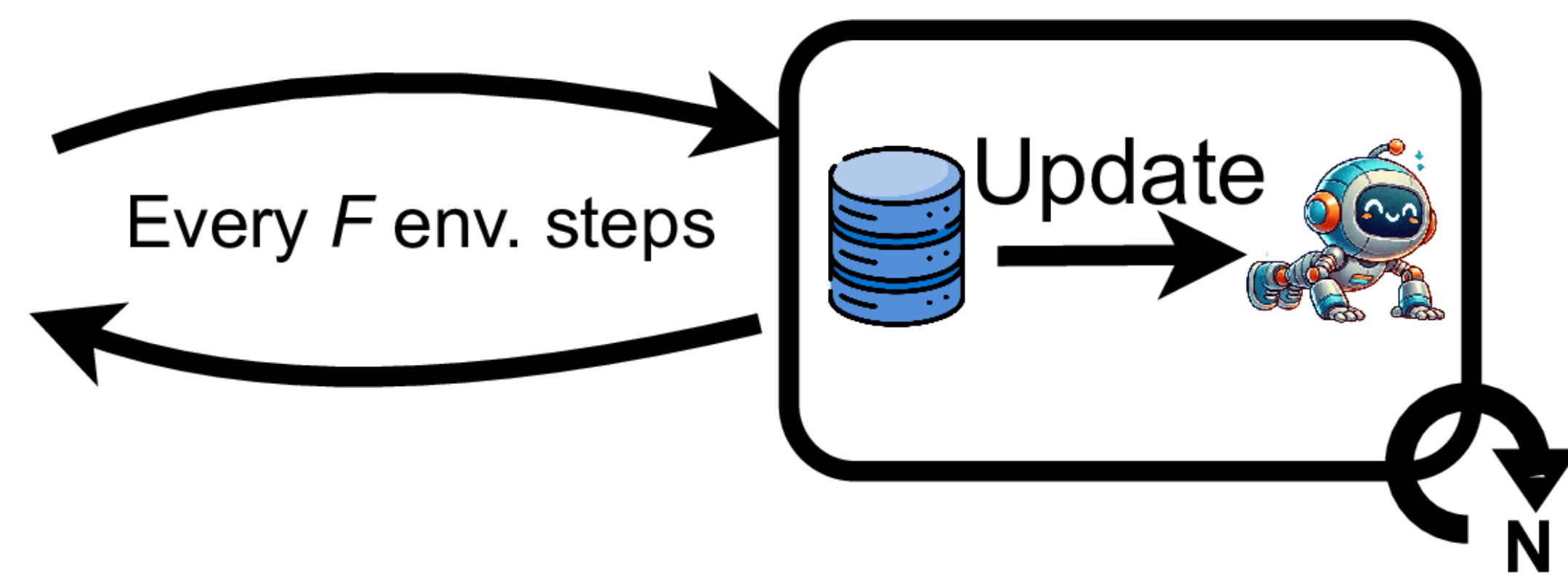


Online low UTD ratio phase



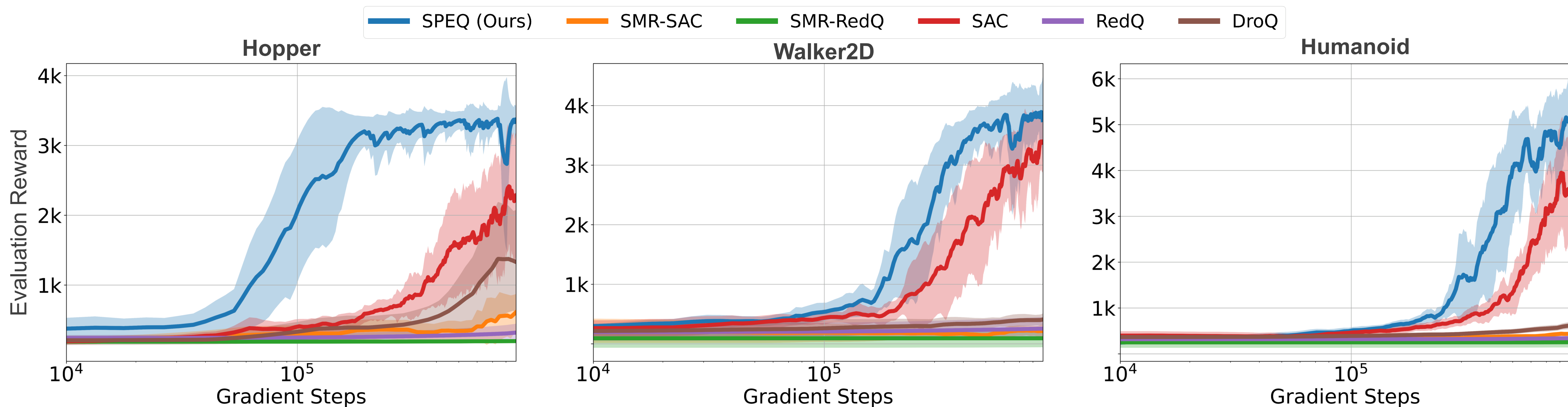
SPEQ

Periodic stabilization phase



Results

SPEQ outperforms high-UTD baselines using up to 99% **fewer gradient updates** and 78% **training time reduction**.



Conclusions



SPEQ alternates two training phases to **reduce waste of gradient updates**.



Achieves SOTA results with lower computational cost, enabling **efficient and scalable RL**.

