Importance Resampling for Off-policy Prediction

Matthew Schlegel, Wesley Chung, Daniel Graves, Jian Qian, Martha White

Motivation

- **Learning off-policy predictions** through general value functions.
- An algorithm with **reduced update variance** and **fewer learning updates**.

Background

(General) Value Function:

$$V(s) \doteq \mathbb{E}_{\pi}\left[\sum_{i=t}^{\infty}\left(\prod_{j=t+1}^{i}\gamma_j\right)C_{i+1} \,|\, S_t = s, A \sim \pi\right]$$

Off-policy Prediction:

- Learn a value function conditioned on a target policy π with data generated from a behavior policy b.

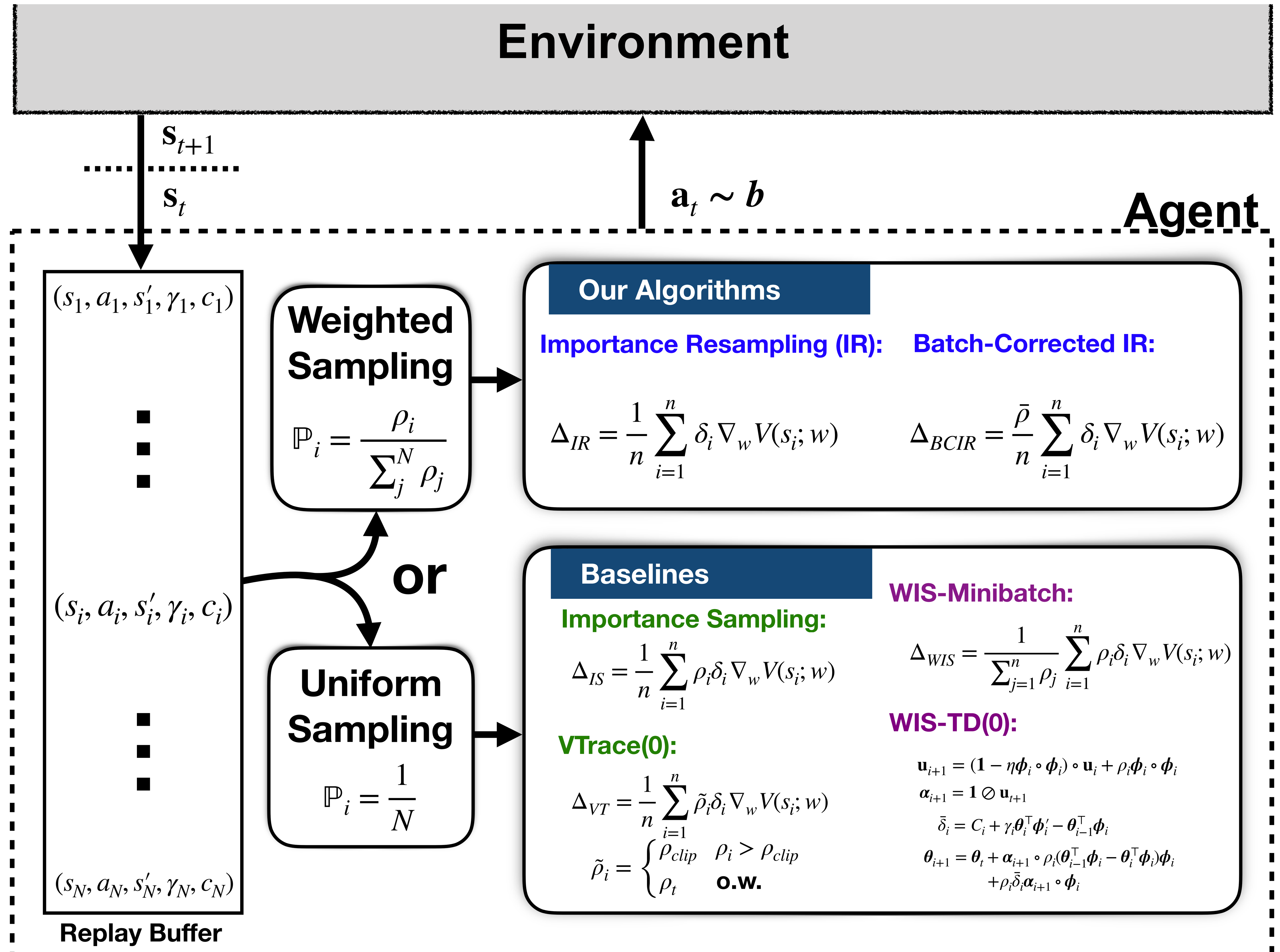
$$\mathbb{E}\left[\Delta_w(A)\,|\,A \sim \pi\right] = \mathbb{E}\left[\rho\Delta_w(A)\,|\,A \sim b\right]$$

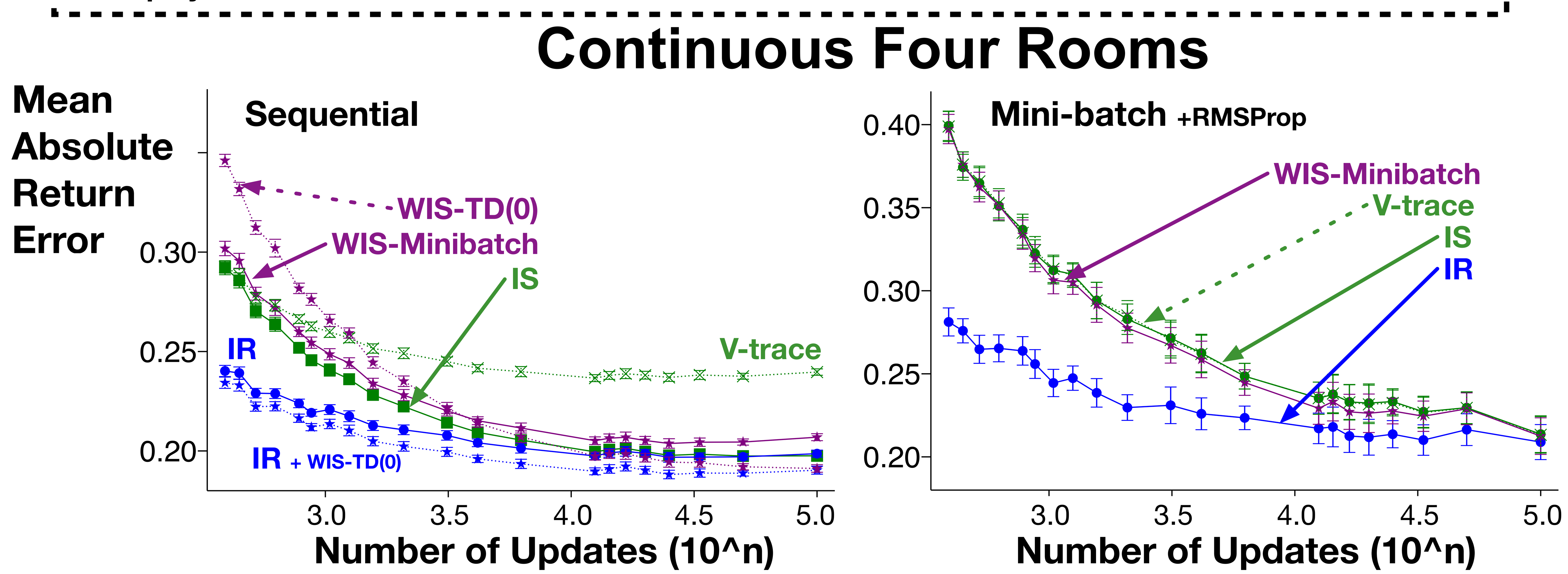
Importance Weights: $\rho_i = \rho(a_i, s_i) = \dfrac{\pi(a_i\,|\,s_i)}{b(a_i\,|\,s_i)}$

RL & AI HUAWEI IVADO

Check out our repo!

https://mkschleg.github.io/Resampling.jl/

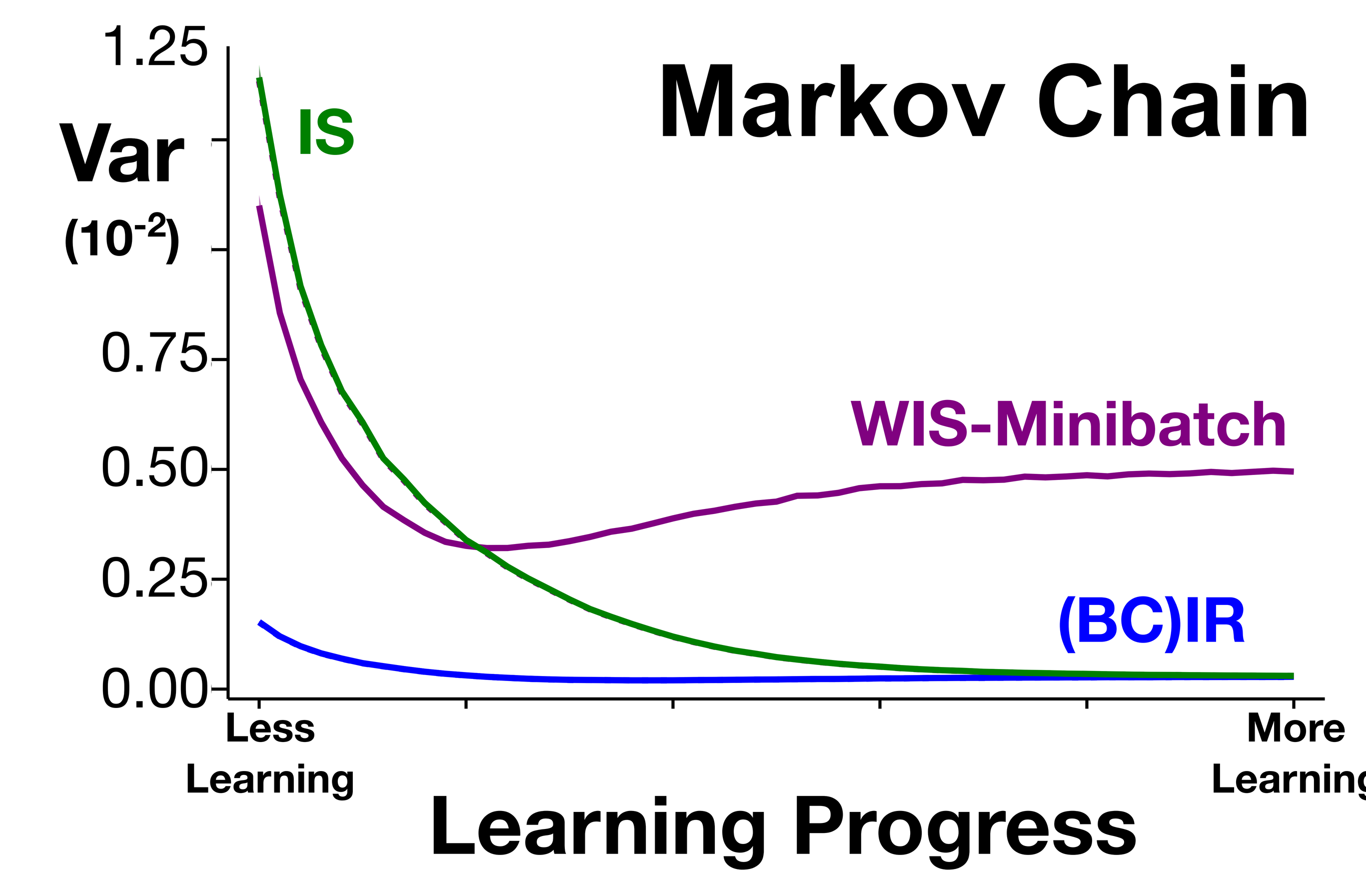




Theoretical Results

- We show BCIR is a **consistent** and **unbiased** estimator of the full batch update in both static buffer and moving buffer scenarios.
- We provide several cases where the variance of IR is less than or equal to that of IS.

IR reduces update variance



IR learns with fewer updates