Importance Resampling for Off-policy Prediction

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

UNIVERSITY OF ALBERTA

Motivation

- Learning many off-policy predictions through general value functions
- Find an approach which corrects for the action distribution and budgets updates more efficiently.

Background

Value Function:

$$V(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A \sim \pi]$$

$$G_t = \sum_{i=t}^{\infty} \left(\prod_{j=t+1}^{i} \gamma_j \right) C_{i+1}$$

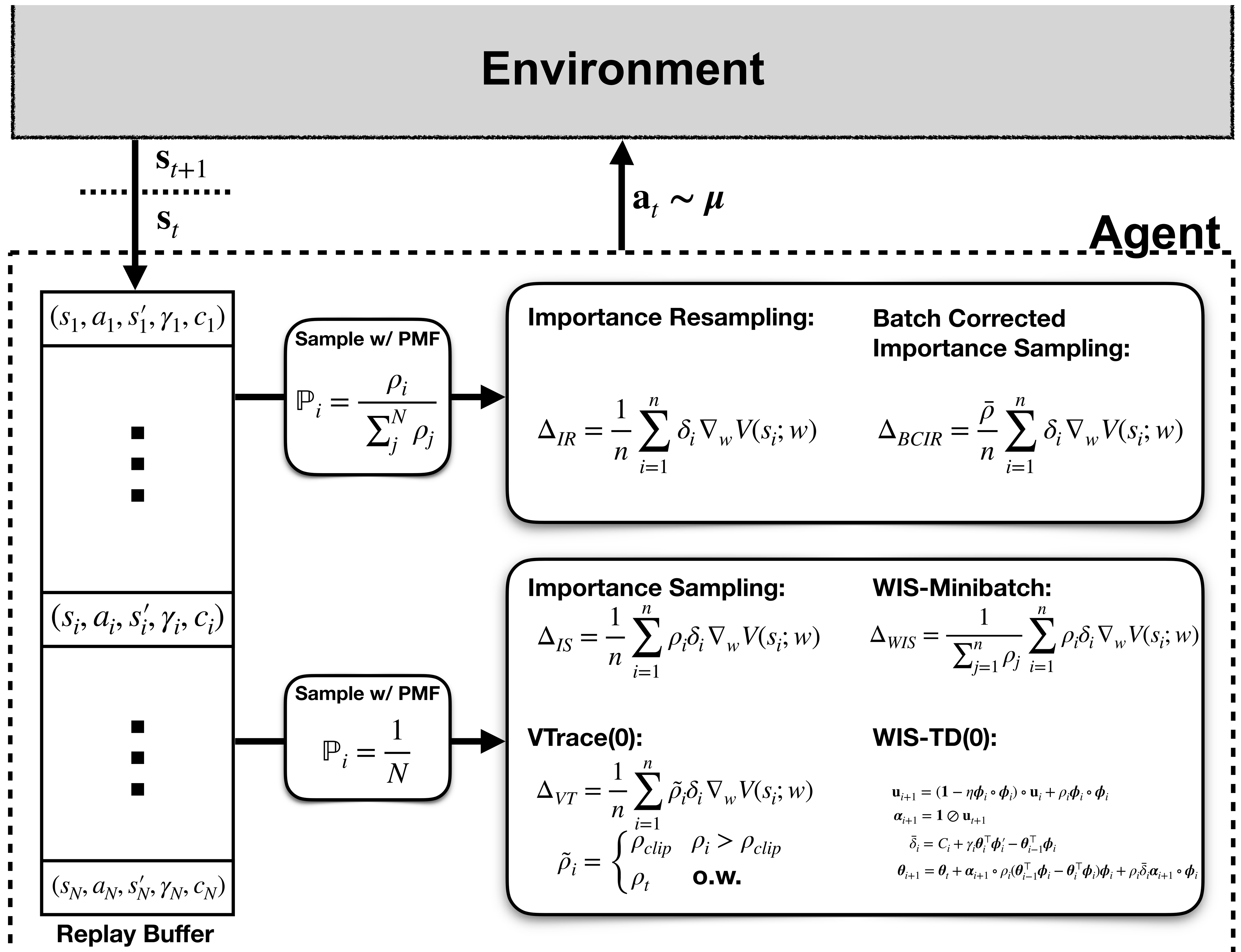
Off-policy Prediction:

Behaviour: $\mu(a \mid s) : A \mapsto \mathbb{R}$

Target: $\pi(a \mid s) : A \mapsto \mathbb{R}$

$$\Delta_{IS} = \mathbb{E}_\mu \left[\rho(A,S)\delta \nabla_w V(S) \right]$$

$$\rho_i = \rho(a_i, s_i) = \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)}$$



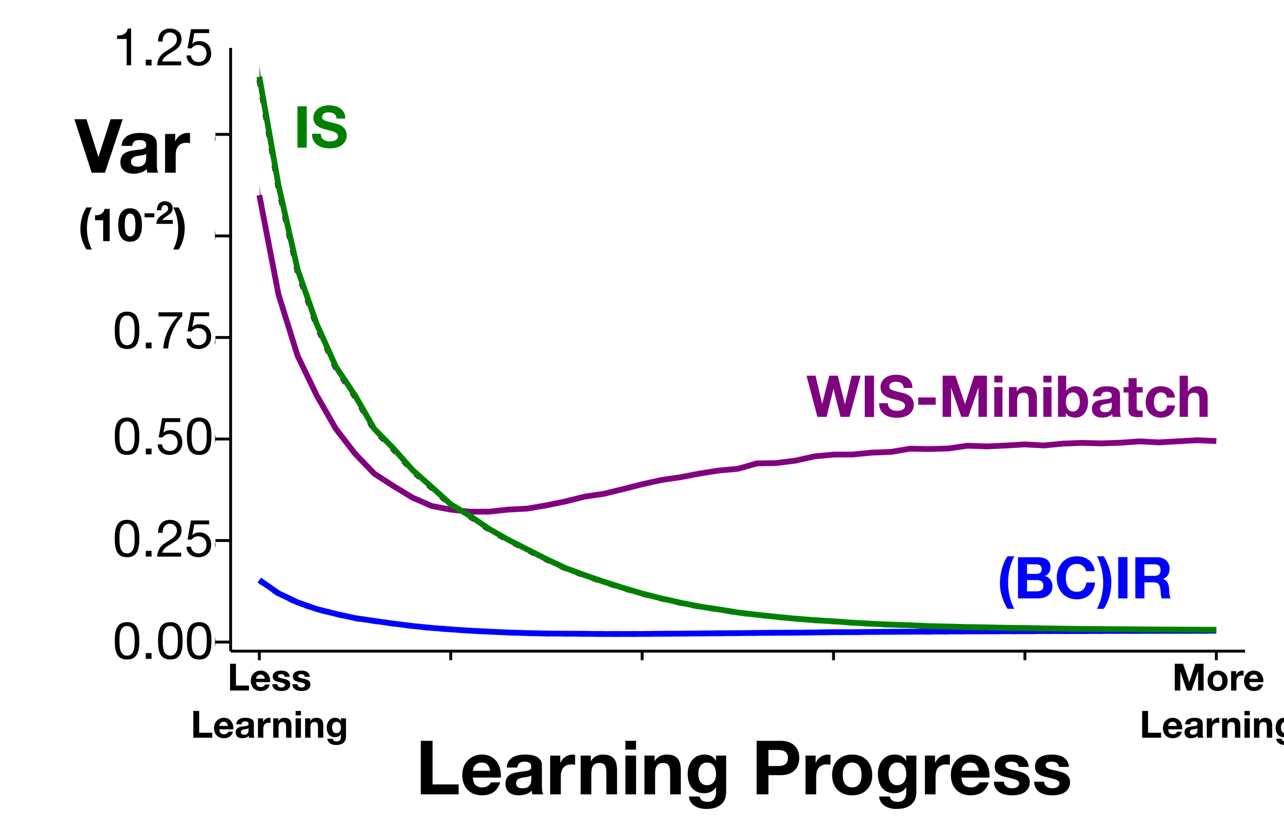
Theoretical Results

- We show BCIR is consistent and unbiased in both the static buffer and moving buffer cases.
- We provide several instances where we expect the variance of IR to be less than that of IS.

Hypothesized Effects

- Update variance reduction compared to importance sampling.
- Faster convergence with respect to the number of updates required.

Markov Chain



Continuous Four Rooms