Importance Resampling:

Conclusions and Future Perspectives

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

Goal

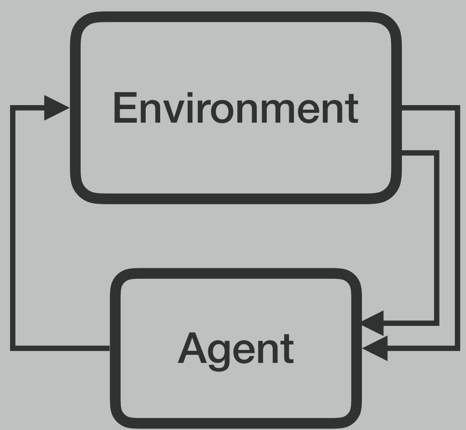
Give an overview of our conclusions from exploring resampling for off-policy prediction.

Outline

- Background
- Reweighting Vs Resampling
- Empirical Results
- Conclusions
- **Future directions and perspectives**

Reweighting

Interact with Environment:



Add $\{\rho_t, S_t, A_t, S_{t+1}\}$ **to** B

Sample Minibatch:

$\{\rho_i, S_i, A_i, S_i'\}$ **with** $Pr\left\{\frac{1}{|B|}\right\}$

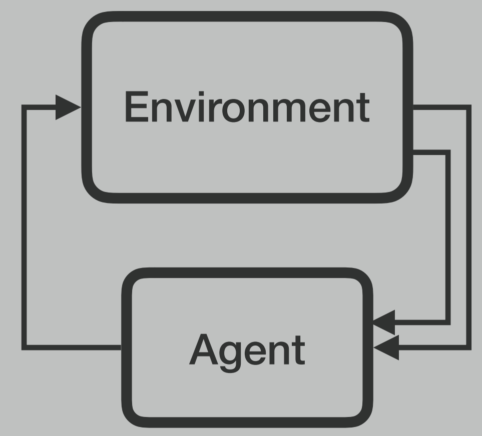
(n times)

Calculate Updates:

$$\Delta_{IS} = \frac{1}{n}\sum_{i=1}^{n} \rho_i \delta_i \nabla_w V(s_i; w)$$

Resampling

Interact with Environment:



Add $\{\rho_t, S_t, A_t, S_{t+1}\}$ **to** B

Update sampling PMF

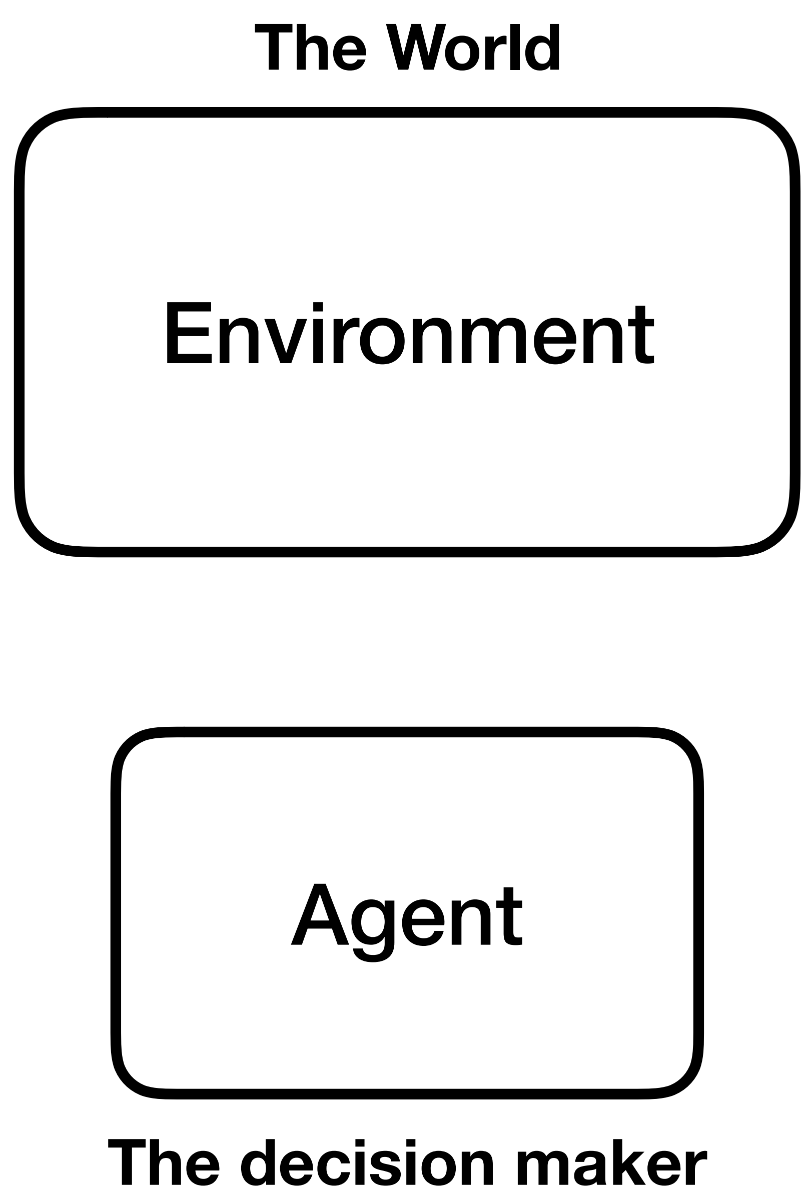
Sample Minibatch:

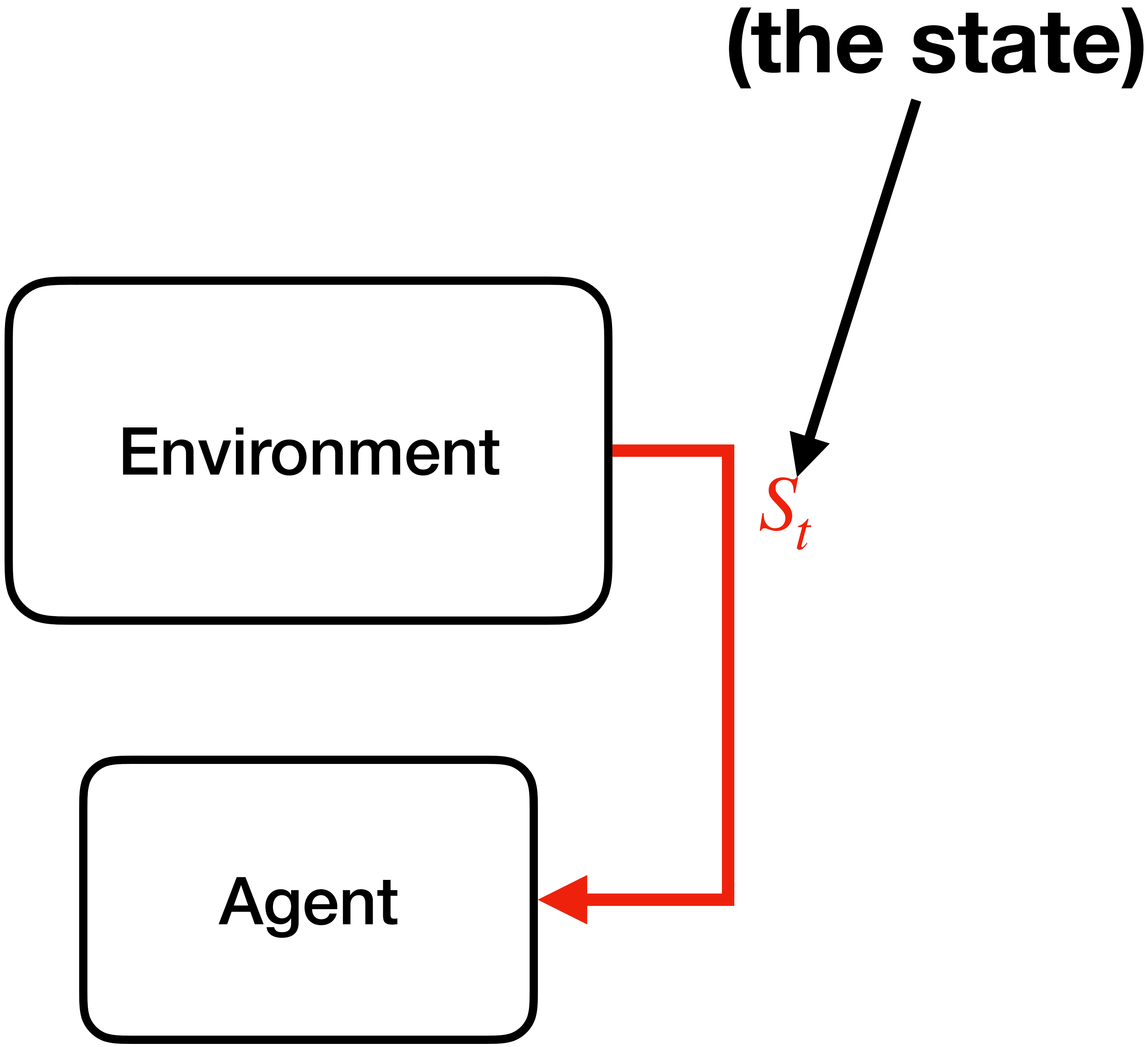
$\{\rho_i, S_i, A_i, S_i'\}$ **with** $Pr\left\{\frac{\rho_i}{\sum_j^{|B|} \rho_j}\right\}$

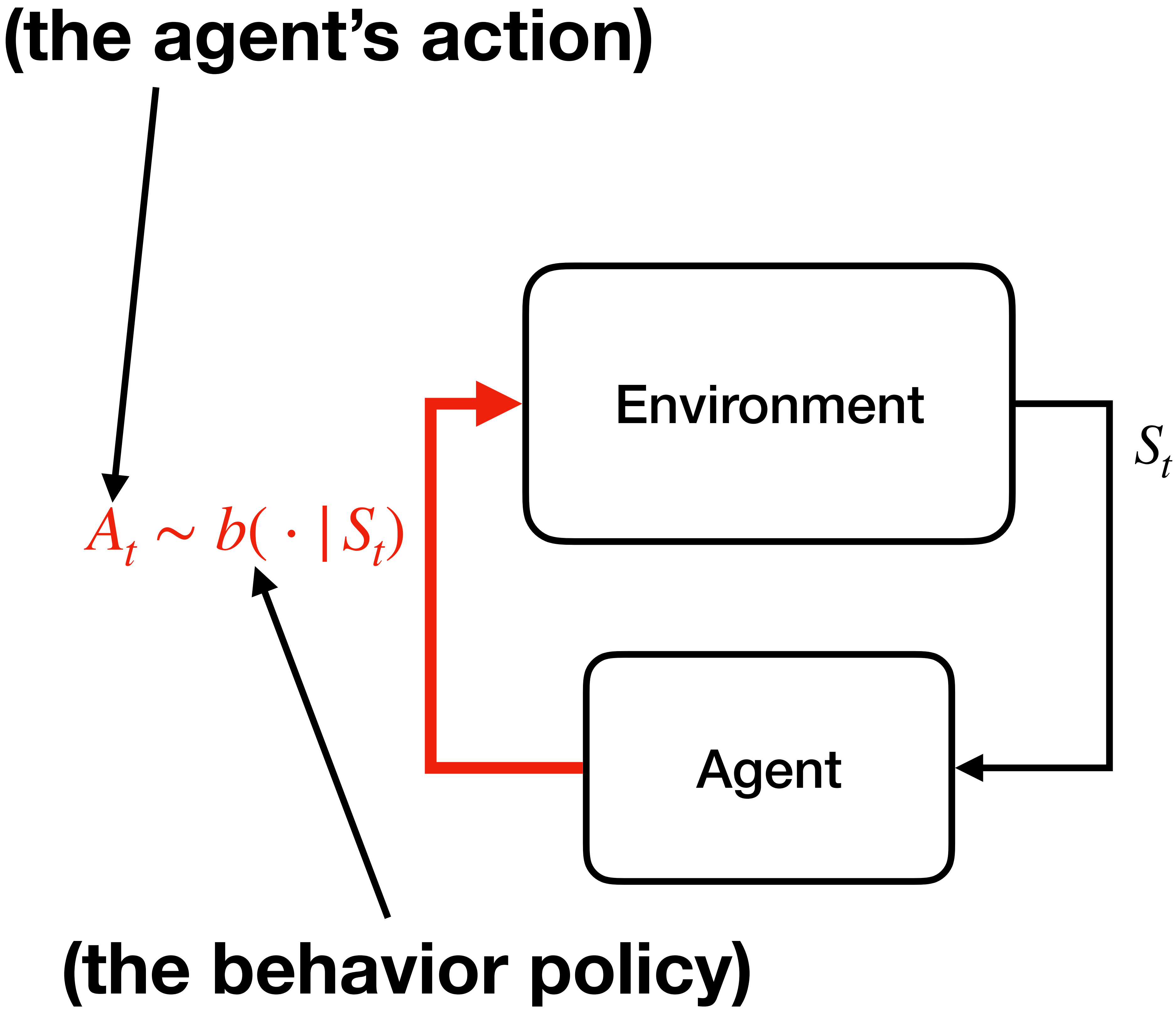
(n times)

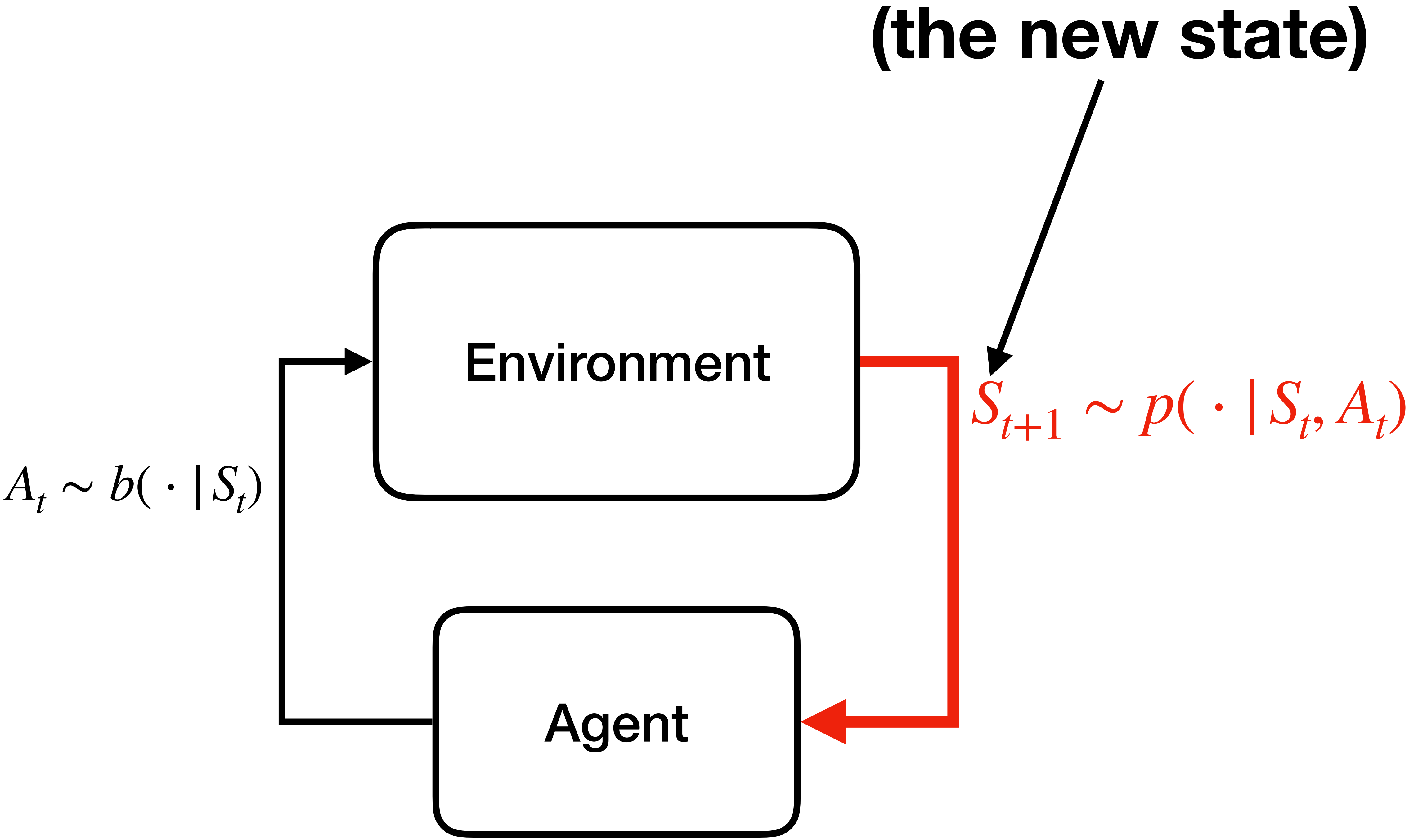
Calculate Updates:

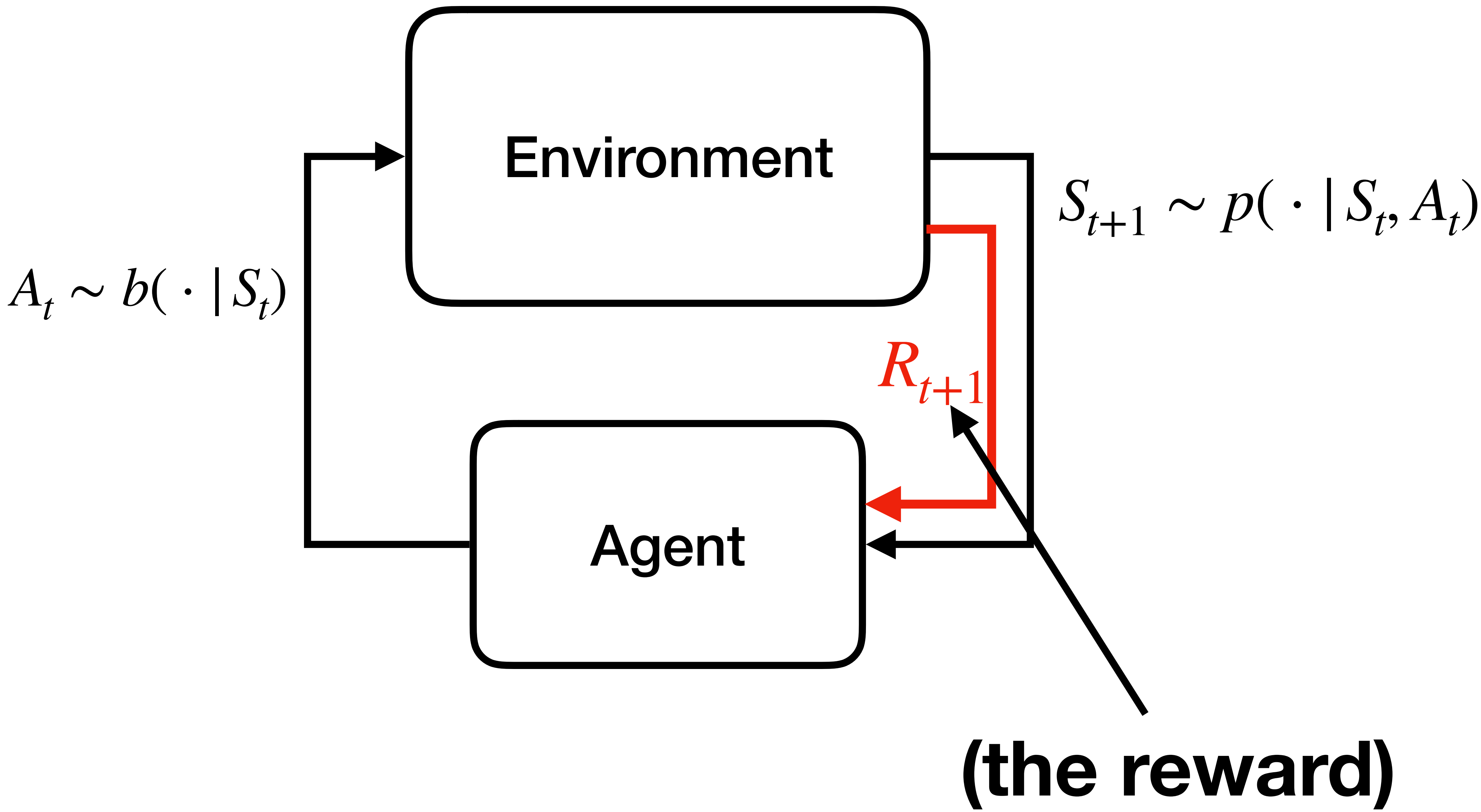
$$\Delta_{IR} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \nabla_w V(s_i; w)$$


The World
Environment
Agent
The decision maker


(the state)
Environment
S_t
Agent


(the agent's action)
$A_t \sim b(\cdot \mid S_t)$
Environment
S_t
Agent
(the behavior policy)


(the new state)
Environment
Agent
$A_t \sim b(\cdot \mid S_t)$
$S_{t+1} \sim p(\cdot \mid S_t, A_t)$



Environment
Agent
$A_t \sim b(\cdot \mid S_t)$
$S_{t+1} \sim p(\cdot \mid S_t, A_t)$
R_{t+1}
(the reward)

The Agent

Agent

Behavior

- Q-learning, Actor Critic, PG…

Predictions (forecasts)

The Agent

Agent

Behavior

- Q-learning, Actor Critic, PG…

Predictions (forecasts)

Buffer of Experience, B

Value Function

$$v(S_t) = \boxed{\mathbb{E}_\pi \left[\sum_{j=t}^{\infty} \left(\prod_{i=t+1}^{h} \gamma(S_i) \right) R_{j+1} \right]}$$

Expected Discounted Return

General Value Function

$$v(S_t) = \mathbb{E}_\pi \left[\sum_{j=t}^{\infty} \left(\prod_{i=t+1}^{h} \gamma(S_i) \right) \boxed{C_{j+1}} \right]$$

Cumulant

Any real-valued signal

General Value Function

$$v(S_t) = \mathbb{E}_{\boxed{\pi}}\left[\sum_{j=t}^{\infty}\left(\prod_{i=t+1}^{h}\gamma(S_i)\right)C_{j+1}\right]$$

Target Policy

$A_{t:\infty} \sim \pi$

Off-policy Learning

Learn about a **target policy** π using data generated from a **behavior policy** b .

Off-policy Learning

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

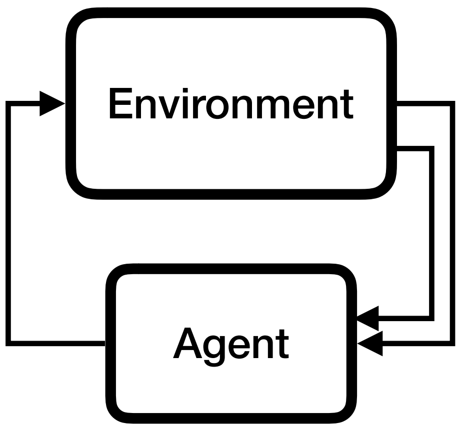
$= \mathbb{E}\left[\boxed{}\Delta_w(A) \,|\, A \sim b\right]$ **Have**

Off-policy Learning

$$
\begin{aligned}
\mathbb{E}\left[\Delta_w(A) \,|\, A \sim \pi\right] &= \sum_{a \in \mathscr{A}} \pi(a)\Delta_w(a) \\
&= \sum_{a \in \mathscr{A}} \pi(a)\frac{b(a)}{b(a)}\Delta_w(a) \\
&= \sum_{a \in \mathscr{A}} \frac{\pi(a)}{b(a)} b(a)\Delta_w(a) \\
&= \mathbb{E}\left[\boxed{\phantom{\frac{\pi(a)}{b(a)}}}\Delta_w(A) \,|\, A \sim b\right]
\end{aligned}
$$

Reweighting

Interact with Environment:



Add $\{\rho_t, S_t, A_t, S_{t+1}\}$ **to** B

Sample Minibatch:

Sample transition $\{\rho_i, S_i, A_i, S_i'\}$ **with** $Pr\left\{\frac{1}{|B|}\right\}$

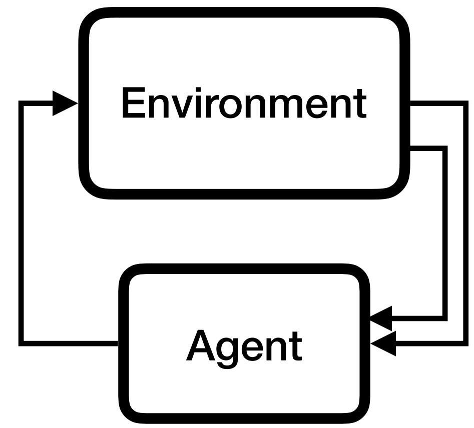
(n times)

Calculate Updates:

$$\Delta_{IS} = \frac{1}{n}\sum_{i=1}^{n} \rho_i \delta_i \nabla_w V(s_i; w)$$

Resampling

Interact with Environment:



Add $\{\rho_t, S_t, A_t, S_{t+1}\}$ **to** B

Update sampling PMF

Sample Minibatch:

Sample transition $\{\rho_i, S_i, A_i, S_i'\}$ **with** $Pr\left\{\frac{\rho_i}{\sum_j^{|B|} \rho_j}\right\}$

(n times)

Calculate Updates:

$$\Delta_{IR} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \nabla_w V(s_i; w)$$

Update Parameters:

$$w' = w - \alpha_t \Delta_*$$

Reweighting

Interact with Environment:

Add $\{\rho_t, S_t, A_t, S_{t+1}\}$ **to** B

Sample Minibatch:

$\{\rho_i, S_i, A_i, S'_i\}$ **with** $Pr\left\{\frac{1}{|B|}\right\}$

(n times)

Calculate Updates:

$$\Delta_{IS} = \frac{1}{n}\sum_{i=1}^{n} \rho_i \delta_i \nabla_w V(s_i; w)$$

Resampling

Interact with Environment:

Add $\{\rho_t, S_t, A_t, S_{t+1}\}$ **to** B

Update sampling PMF

Sample Minibatch:

$\{\rho_i, S_i, A_i, S'_i\}$ **with** $Pr\left\{\frac{\rho_i}{\sum_j^{|B|} \rho_j}\right\}$

(n times)

Calculate Updates:

$$\Delta_{IR} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \nabla_w V(s_i; w)$$

Off-policy Learning

With a buffer of experience

Importance Sampling (IS):

$$\Delta_{IS} = \frac{1}{n} \sum_{i=1}^{n} \rho_i \delta_i \nabla_w V(s_i; w)$$

Importance Resampling (IR):

$$\Delta_{IR} = \frac{1}{n} \sum_{i=1}^{n} \delta_i \nabla_w V(s_i; w)$$

WIS-Minibatch:

$$\Delta_{WIS} = \frac{1}{\sum_{j=1}^{n} \rho_j} \sum_{i=1}^{n} \rho_i \delta_i \nabla_w V(s_i; w)$$

VTrace(0):

$$\bar{\rho}_i = \begin{cases} \rho_{clip} & \rho_i > \rho_{clip} \\ \rho_t & \textbf{o.w.} \end{cases}$$

$$\Delta_{VTrace} = \frac{1}{n} \sum_{i=1}^{n} \bar{\rho}_i \delta_i \nabla_w V(s_i; w)$$

Off-policy Learning

With a buffer of experience

Reweighting

Importance Sampling (IS)

VTrace(0)

WIS-Minibatch

Resampling

Importance Resampling (IR)

Hypothesized Empirical Benefits

- IR reduces the **update variance** as compared with IS.

- IR can **update less to learn more** (sample efficiency).

Variance in Off-policy Prediction

Update Variance:

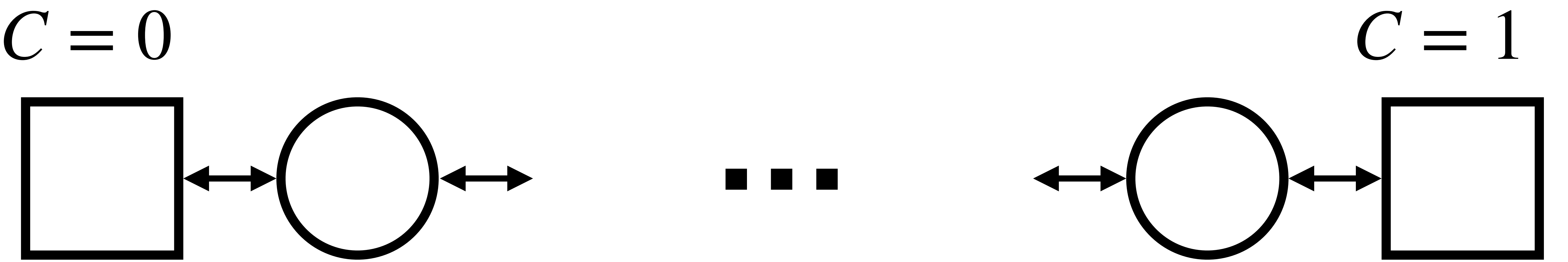
$$Var\left\{\Delta_{IS}\right\} = Var\left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \rho_i \delta_i \nabla_w V(s_i; w) \right\|_1 \right\}$$

Benefits of reduced update variance:

- Reduced sensitivity to learning rate.
- Faster learning

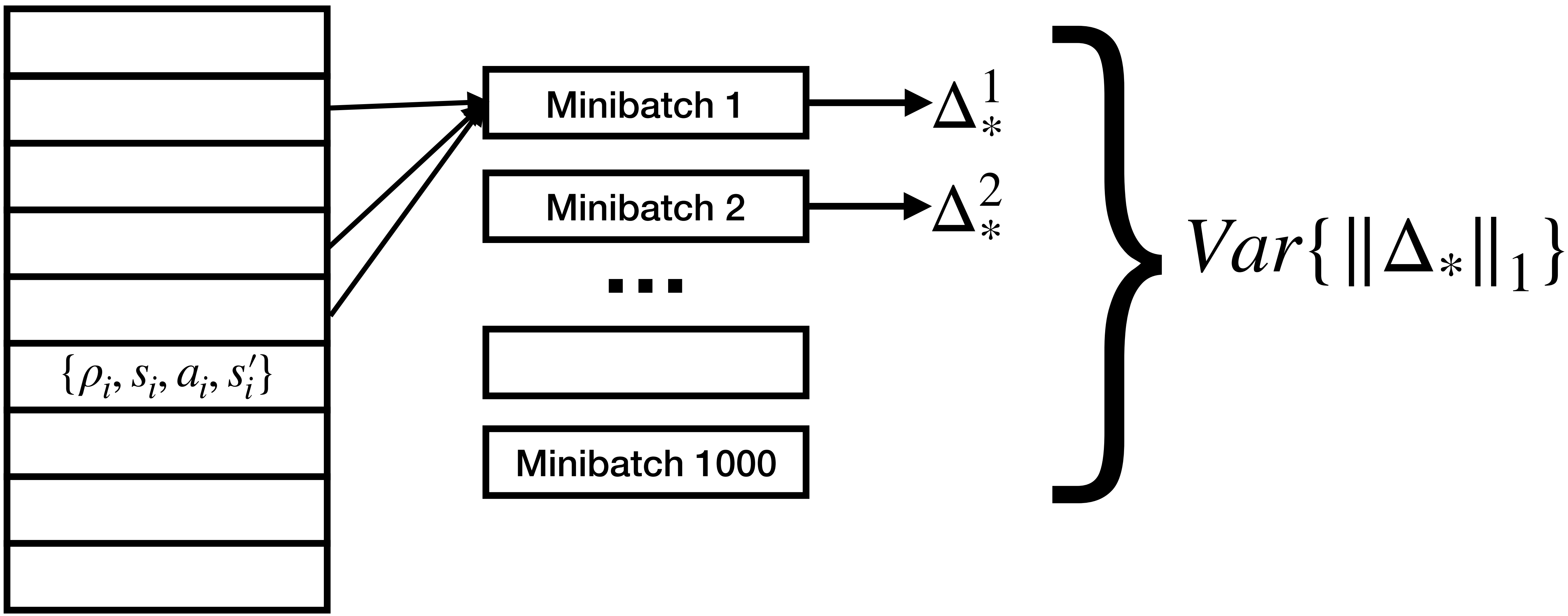
Empirical Results

Markov Chain



Markov Chain

Estimating the Update Variance:



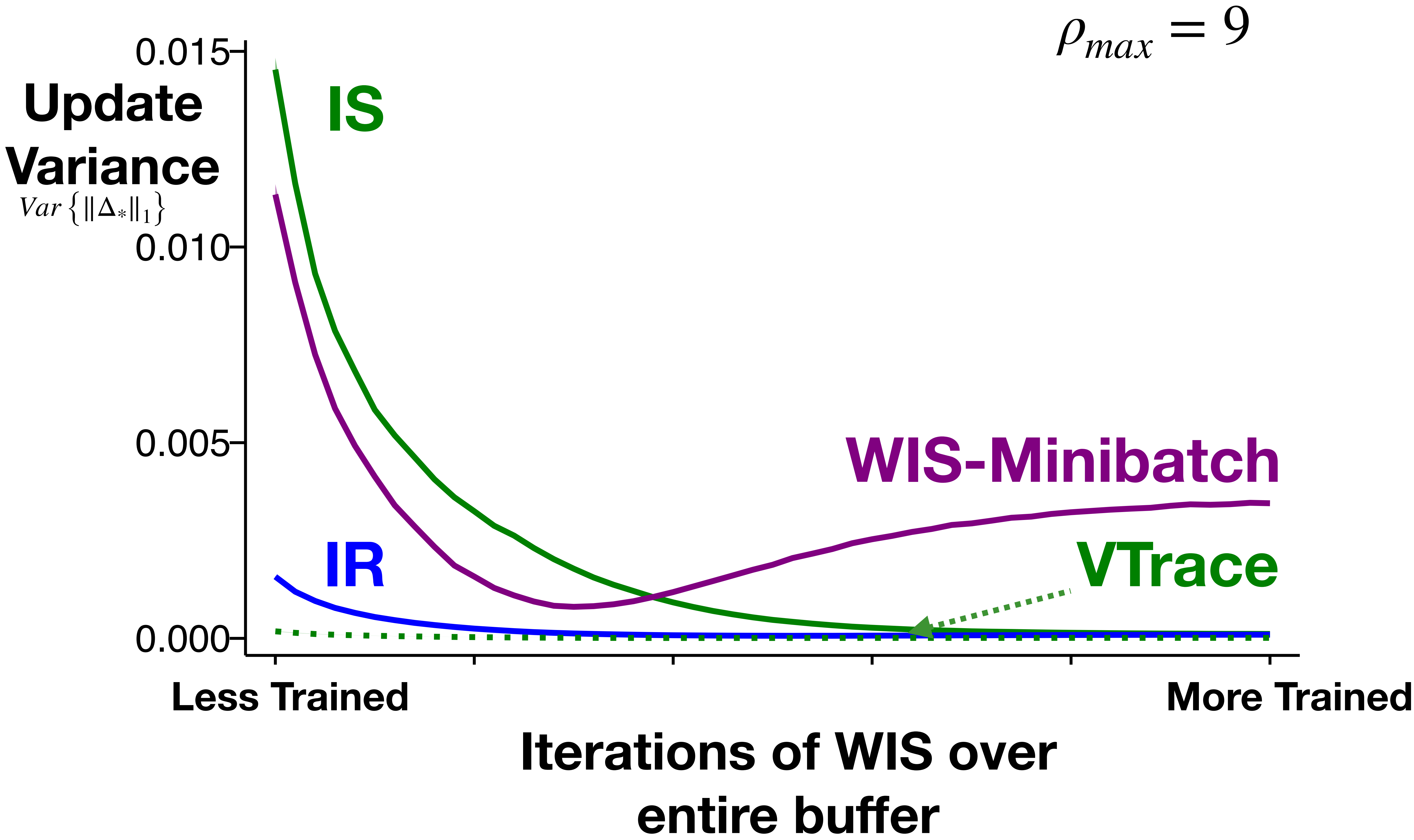
Behavior:

$$b(a\,|\,s) = \begin{cases} 0.9 & \textbf{if } a = left \\ 0.1 & \textbf{if } a = right \end{cases}$$

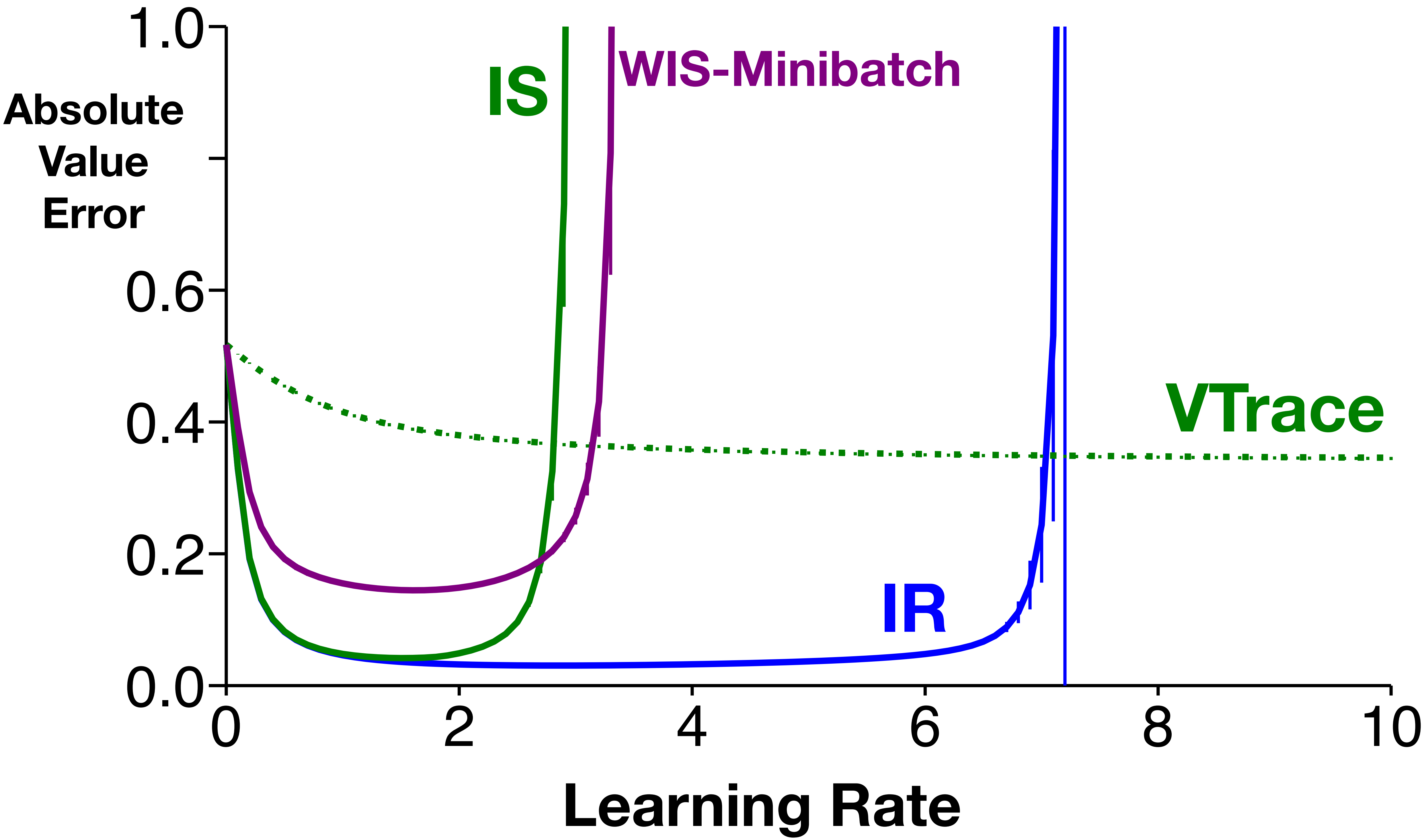
Target:

$$\pi(a\,|\,s) = \begin{cases} 0.1 & \textbf{if } a = left \\ 0.9 & \textbf{if } a = right \end{cases}$$

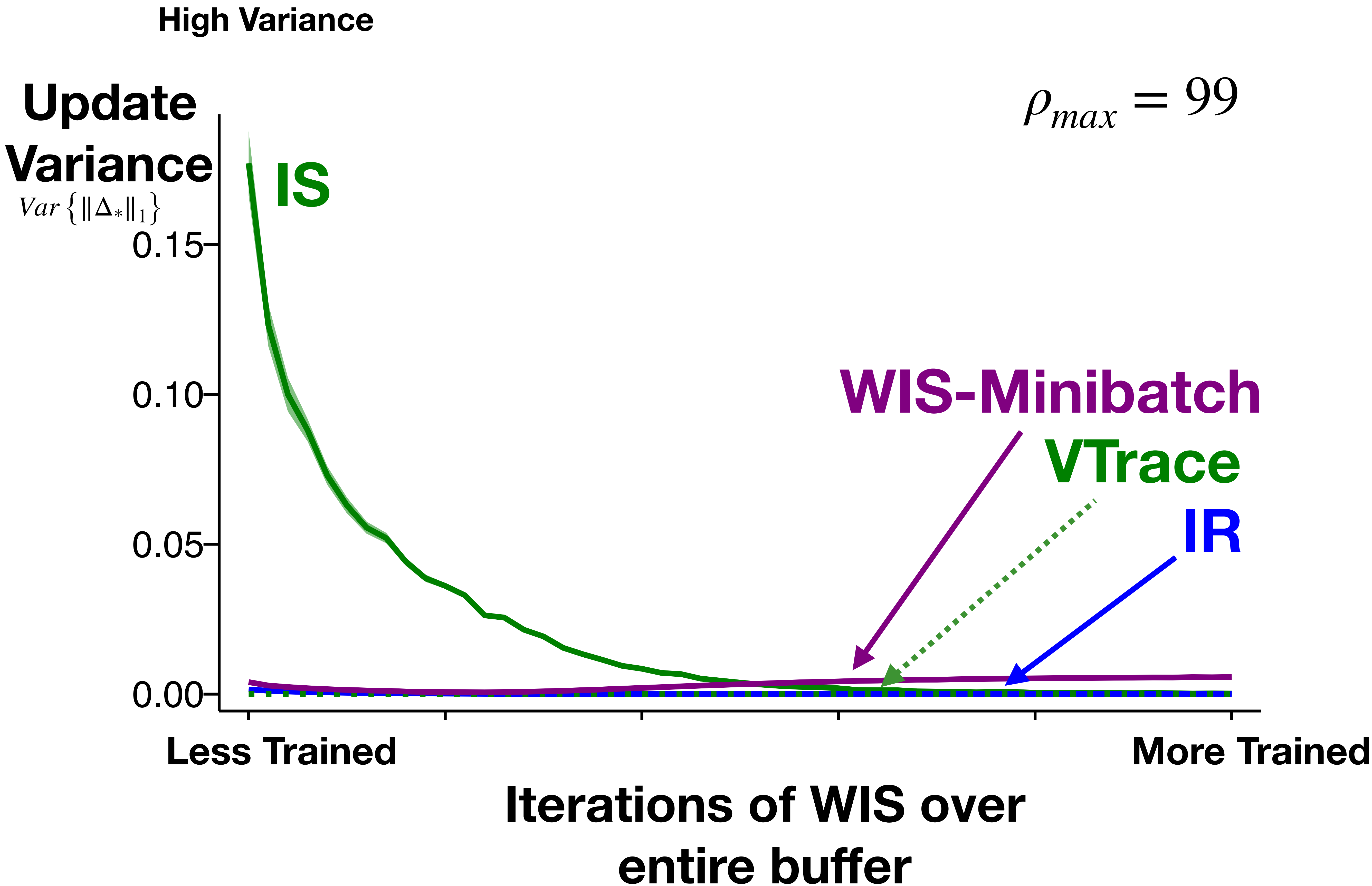
Markov Chain - Update Variance



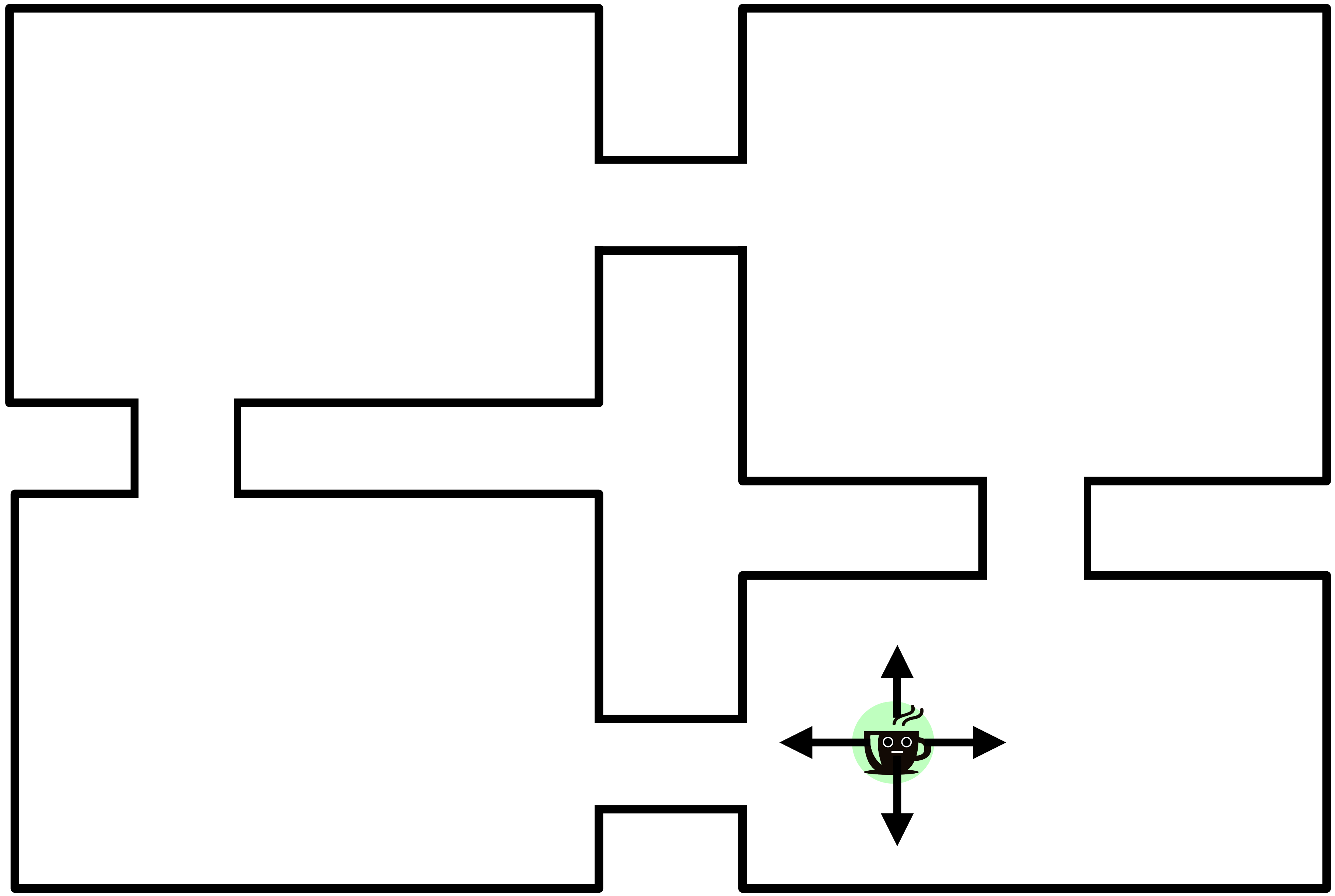
Markov Chain - Learning Rate Sensitivity



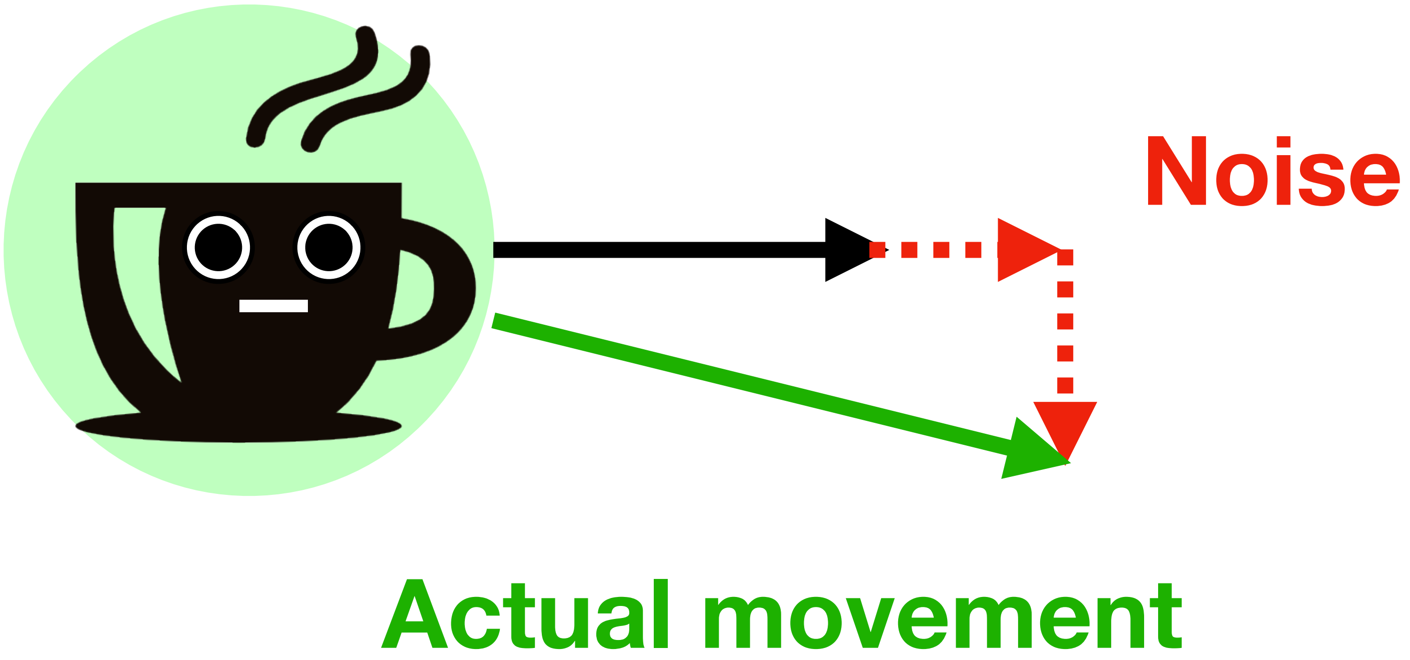
Markov Chain - Update Variance



Continuous Four Rooms



Continuous Four Rooms



Continuous Four Rooms

Evaluation:

- Sampled 1000 states from the stationary distribution of the behavior policy
- Estimated returns with 100 Monte Carlo rollouts

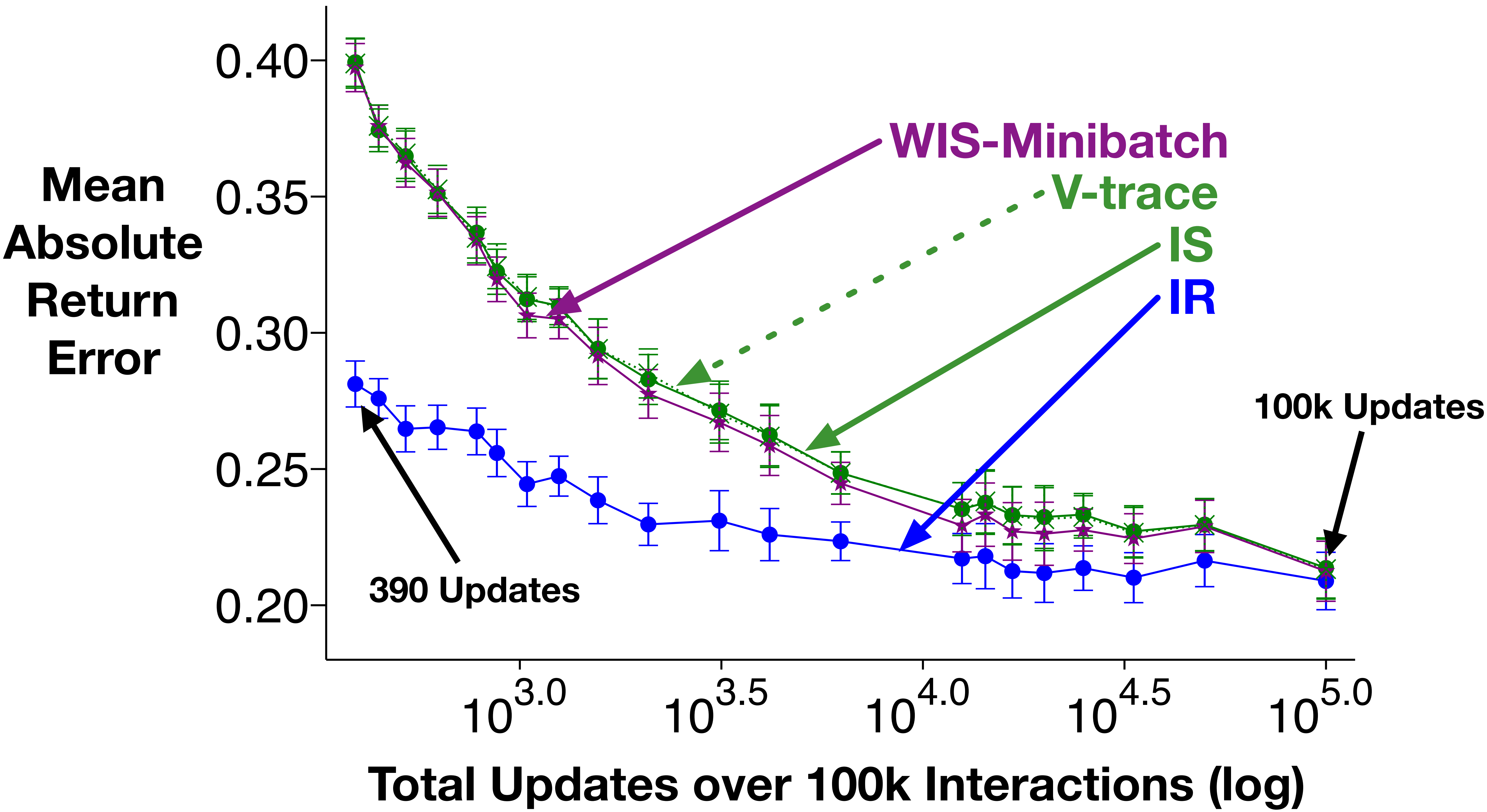
Behavior:

$$b(\cdot \mid s) = 0.25$$

Target:

$$\pi_1(a \mid s) = \begin{cases} 1 & \textbf{if } a = \textbf{down} \\ 0 & \textbf{o.w.} \end{cases}$$

Cont. Four Rooms - Total Updates



Conclusions

1. Resampling can have **lower variant updates** as compared to importance sampling.

2. Resampling generally needs **fewer updates** to reach comparable performance to importance sampling.

3. Resampling and importance sampling **perform comparably when many samples are used.**

Buffer of experience

~~Should we update all predictions at every interaction?~~

Do we have to update all predictions at every interaction?

Maybe not?

@mattschleg
mkschleg
github.com/mkschleg
coffeesideai.com

Questions?

More Experiments!

Weird behavior of induced bias!!

Theory!

https://arxiv.org/pdf/1906.04328.pdf

HUAWEI
RL & AI
amii
UNIVERSITY OF ALBERTA