Importance Resampling:

Conclusions and Future Perspectives

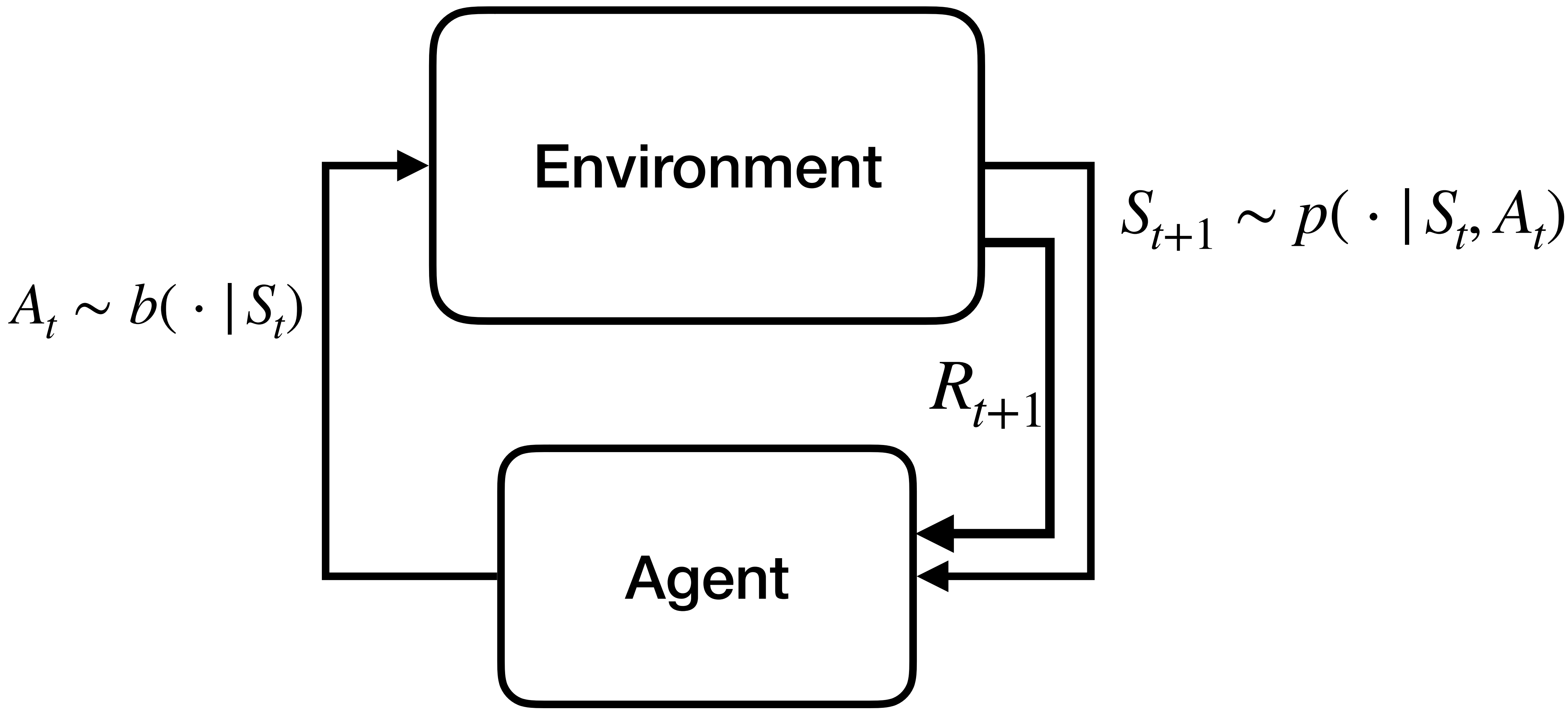
Matthew Schlegel, Wes Chung, Jian Qian,
Daniel Graves, and Martha White

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**


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

Value Function

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

Expected Discounted Return

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

Behavior Policy: $A_{t:\infty} \sim \mu$

Discount: $\gamma(s_t) \in [0,1]$

Reward/Cumulant: $r(s_t) \in \mathbb{R}$

Off-policy Learning

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

Off-policy Learning

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

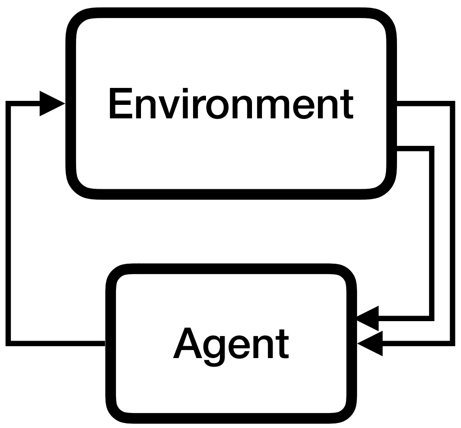
$= \mathbb{E}\left[\boxed{}\Delta_w(A) \,|\, A \sim b\right]$ **Have**

Off-policy Learning

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

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** $Pr\left\{\frac{1}{|B|}\right\}$

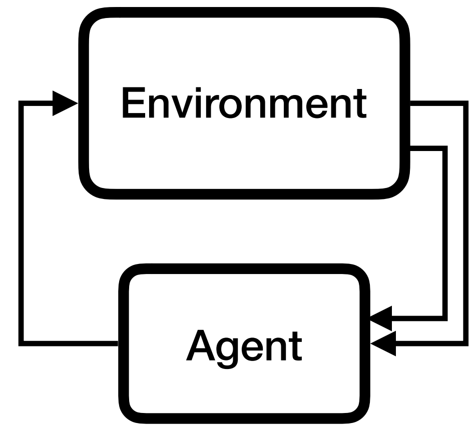
(n times)

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)

Off-policy Learning

WIS-TD(0)

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:

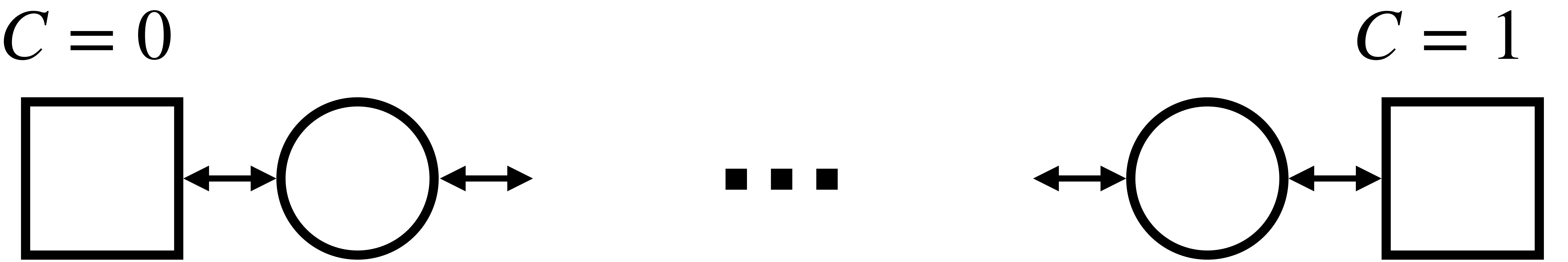
$$Var\left\{\Delta_{IS}\right\} = Var\left\{ \left\| \frac{1}{n}\sum_{i=1}^{n}\rho_i\delta_i\nabla_w V(s_i;w) \right\|_1 \right\}$$

Benefits of reduced update variance:

- Reduced sensitivity to learning rate.
- Faster learning

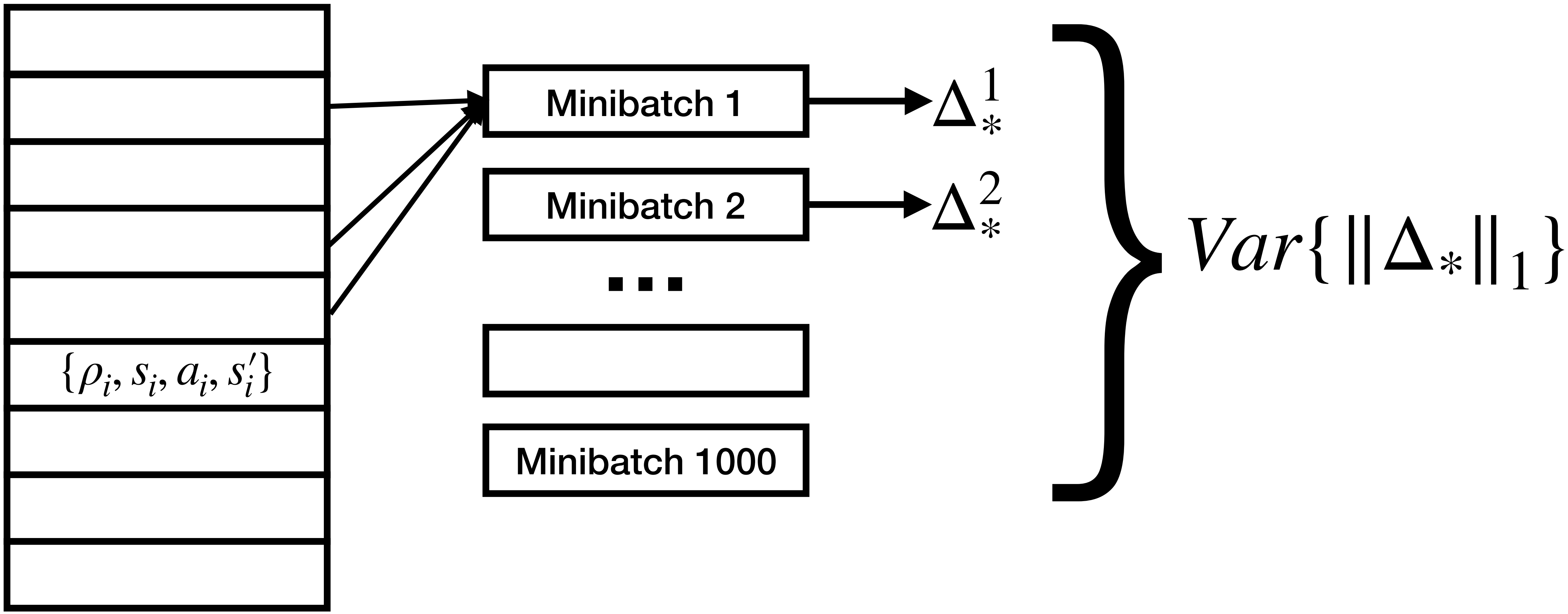
Empirical Results

Markov Chain



Markov Chain

Estimating the Update Variance:



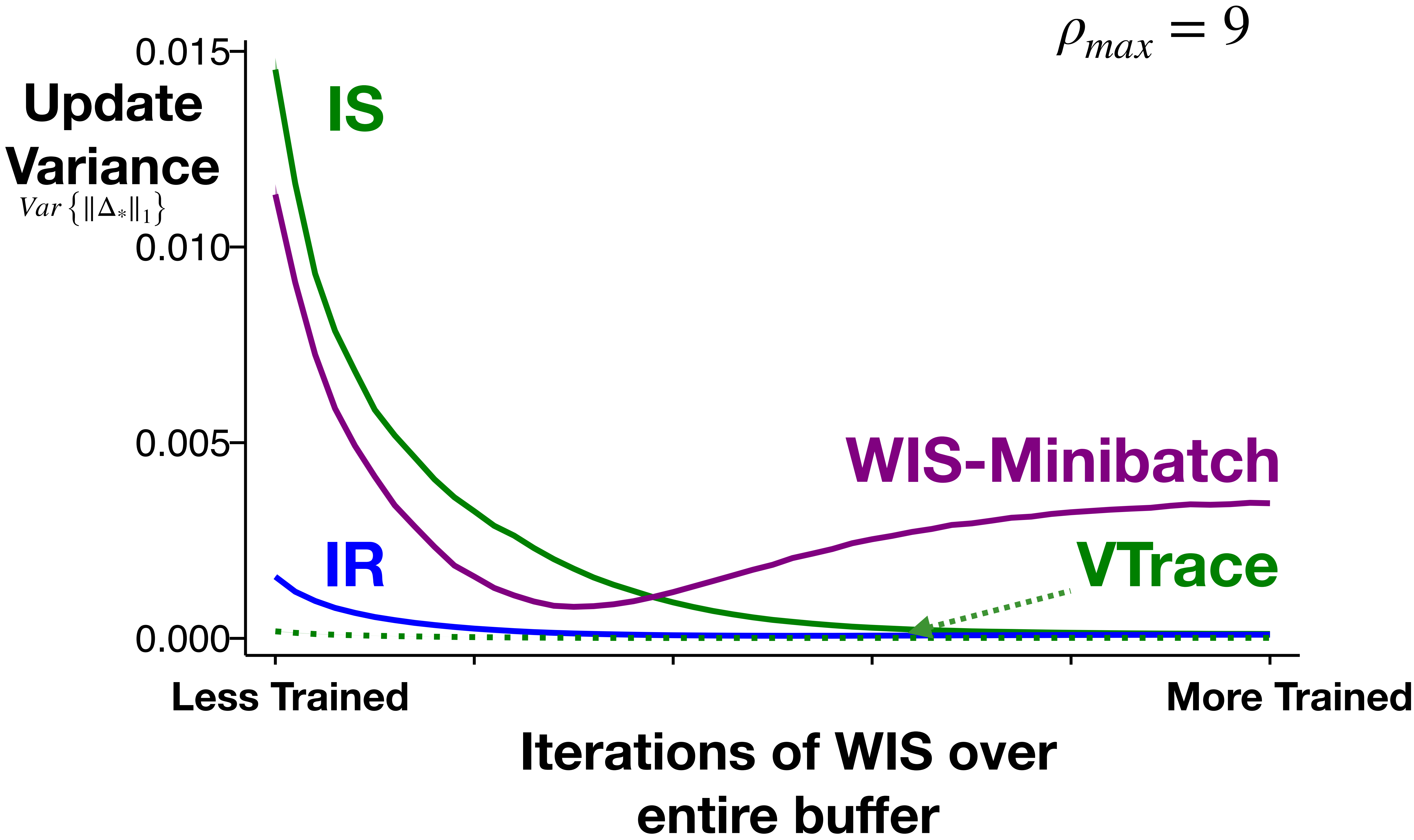
Behavior:

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

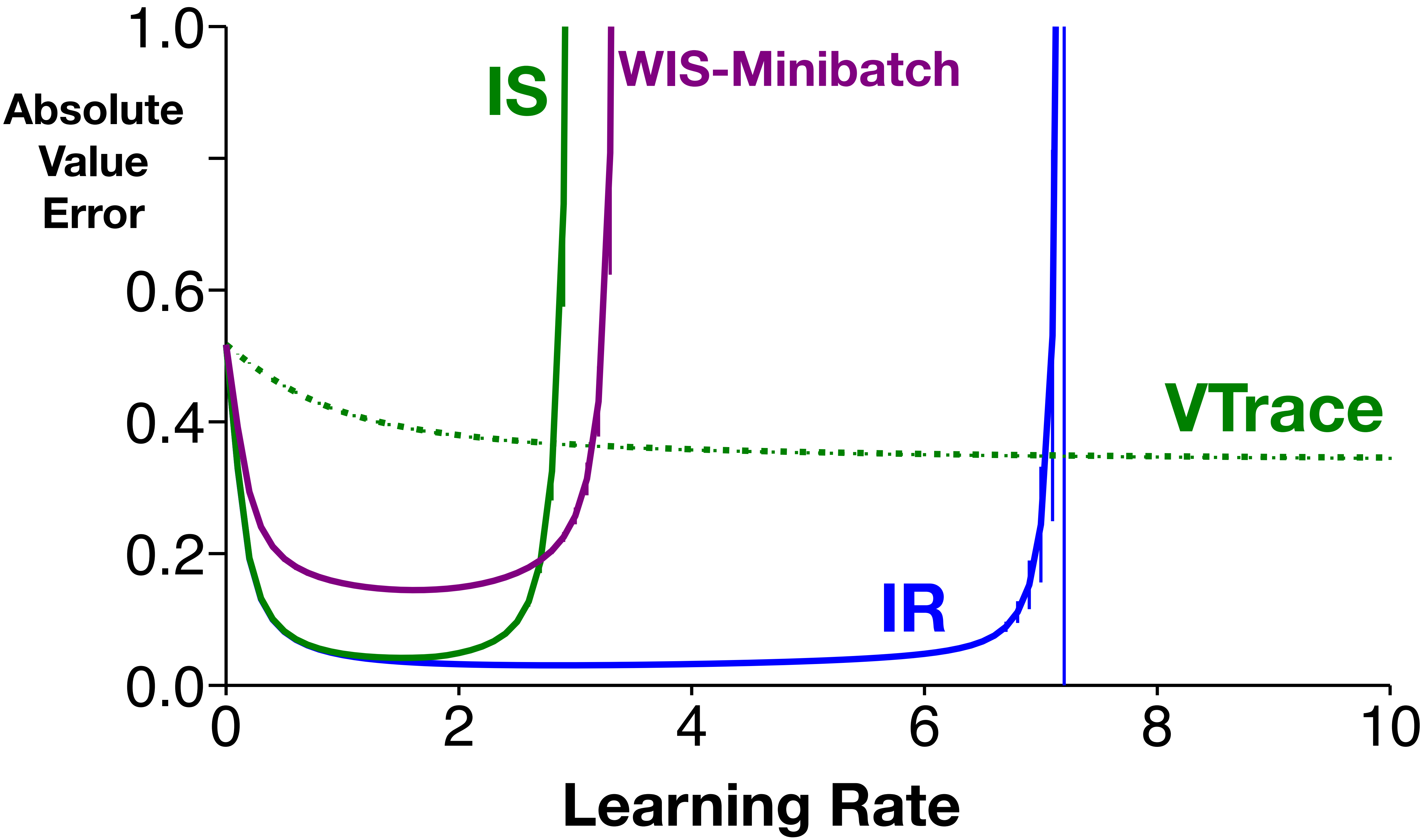
Target:

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

Markov Chain - Update Variance

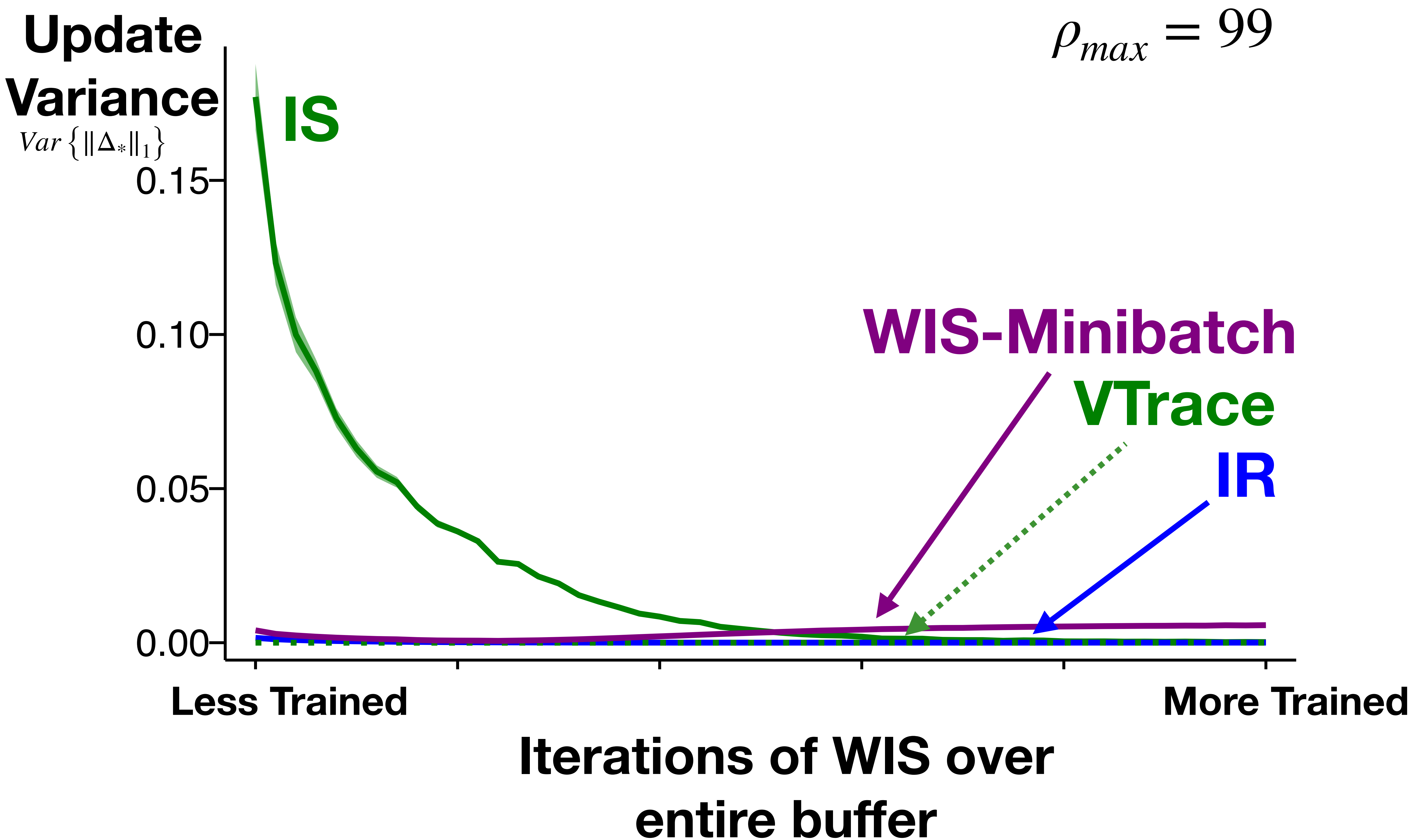


Markov Chain - Learning Rate Sensitivity

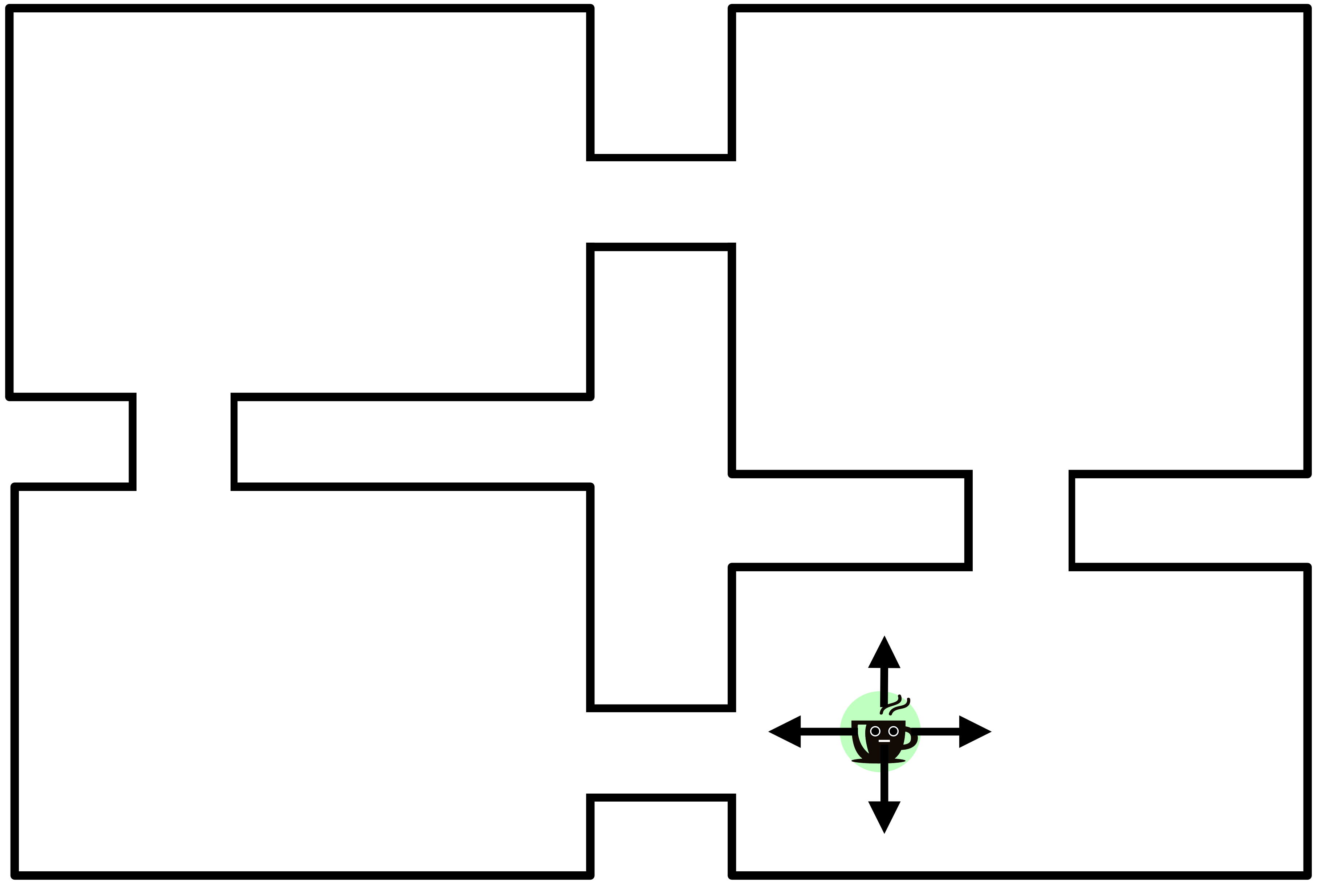


Markov Chain - Update Variance

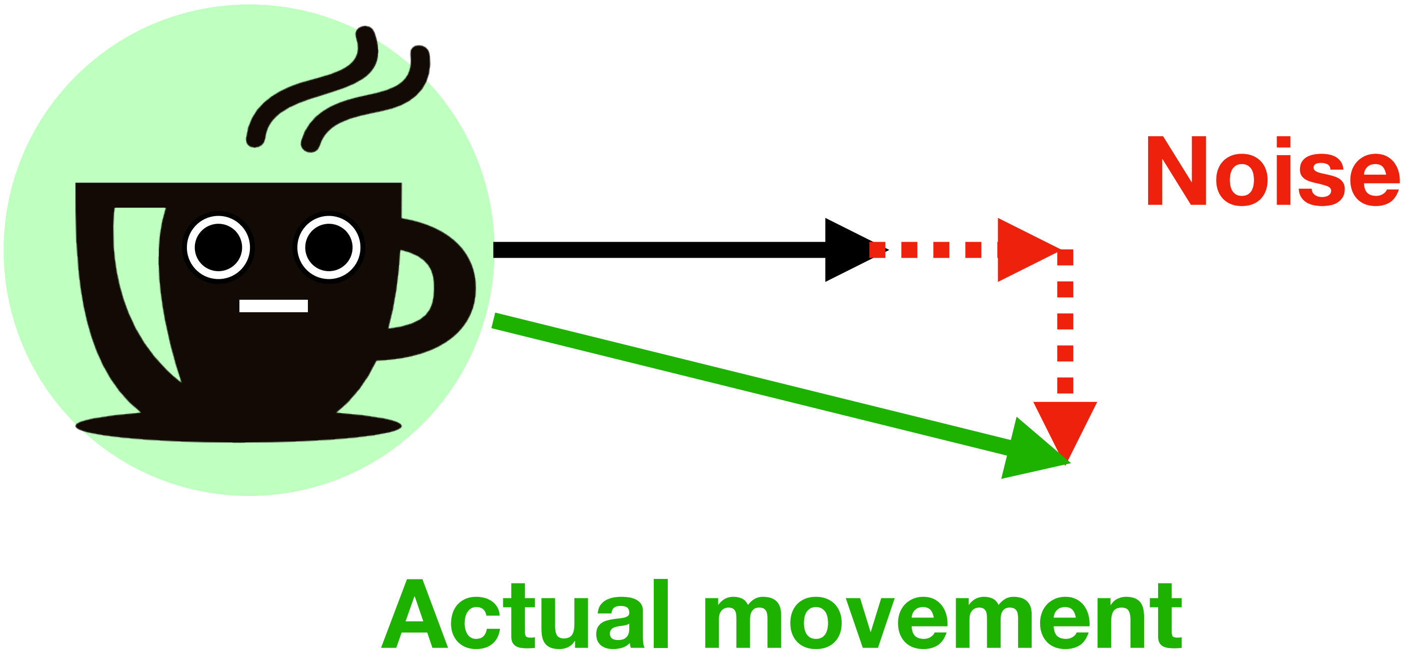
High Variance



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

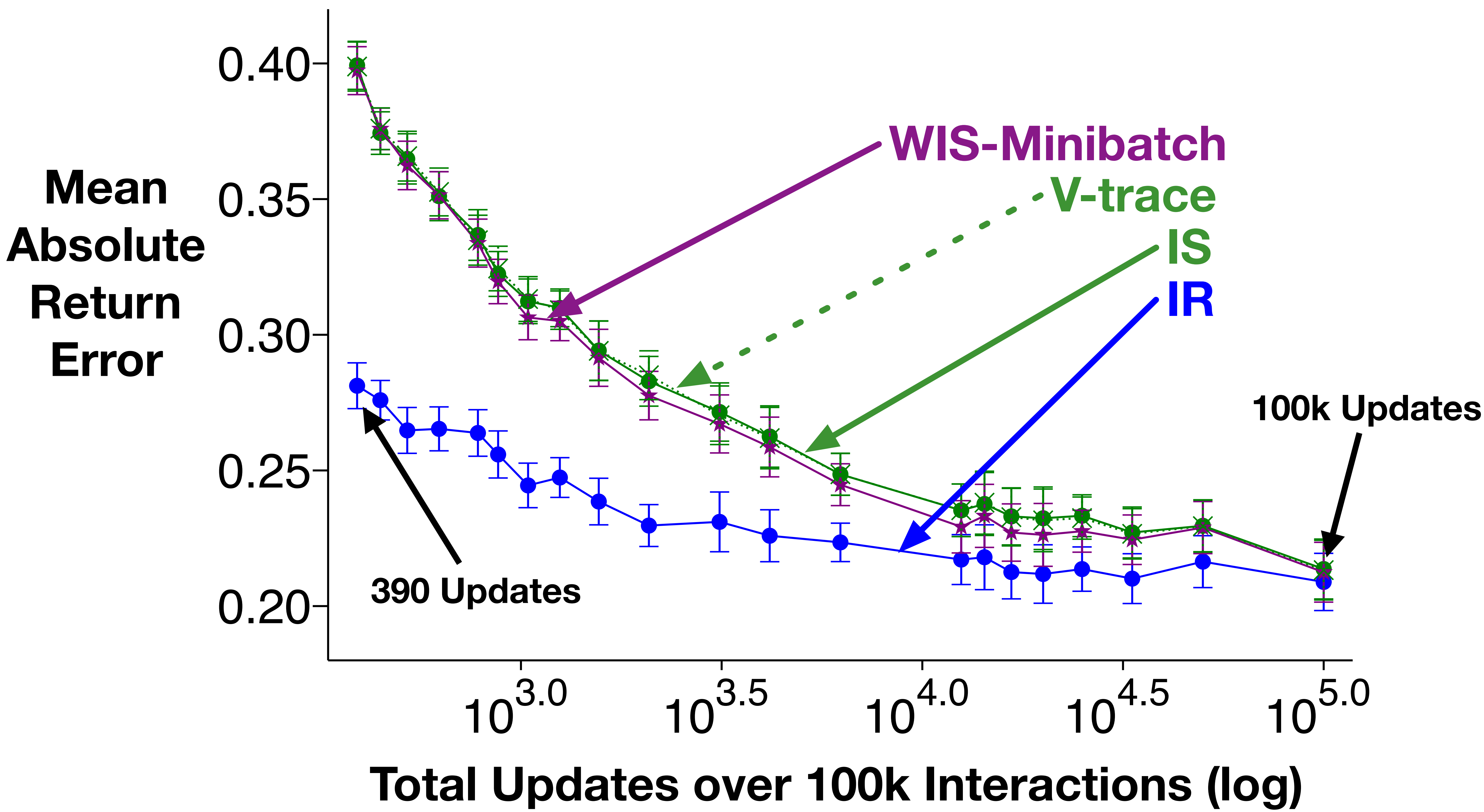
Behavior:

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

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.**

@mattschleg
mkschleg
github.com/mkschleg
coffeesideai.com

Questions?

More Experiments!

Weird behavior of induced bias!!

Theory!

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

Theory

Theoretical Properties of IR

- Biased and Consistent (with a small correction term)
- Consistent (with the correction term) under a changing buffer of experience.
- Under some conditions, guaranteed equal or lower variance compared to IS.

Bias and Consistency

Bias

Theorem 3.1. *[Bias for a fixed buffer of size n] Assume a buffer B of n transitions is sampled i.i.d., according to $d_\mu(s)\mu(a|s)\mathrm{P}(s'|s,a)$. Let $X_{\mathrm{WIS}^*} \stackrel{\text{def}}{=} \sum_{i=1}^{n} \frac{\rho_i}{\sum_{j=1}^{n} \rho_j} \Delta_i$ be the WIS-Optimal estimator of the update. Then,*

$$\mathbb{E}[X_{\mathrm{IR}}] = \mathbb{E}[X_{\mathrm{WIS}^*}]$$

and so the bias of X_{IR} is proportional to

$$\mathrm{Bias}(X_{\mathrm{IR}}) = \mathbb{E}[X_{\mathrm{IR}}] - \mathbb{E}_\pi[\Delta] \propto \frac{1}{n}(\mathbb{E}_\pi[\Delta]\sigma_\rho^2 - \sigma_{\rho,\Delta}\sigma_\rho\sigma_\Delta) \tag{1}$$

Consistency

Theorem 3.2. *Let $B_i = \{X_{i-n+1}, ..., X_i\}$ be the buffer of the most recent n transitions sampled by time i, i.i.d. as specified in Assumption 1. Let $X_{\mathrm{BC}}^{(i)}$ be the bias-corrected IR estimator, with k samples from buffer B_i. Define the sliding-window estimator $X_t \stackrel{\text{def}}{=} \frac{1}{t}\sum_{i=1}^{t} X_{\mathrm{BC}}^{(i)}$. Assume there exists a $c > 0$ such that $\mathrm{Var}(X_{\mathrm{BC}}^{(i)}) \leq c\ \forall i$. Then, as $t \to \infty$, X_t converges in probability to $\mathbb{E}_\pi[\Delta]$.*

Variance

Theorem 3.3. *Assume that, for a given buffer B, $\|\Delta_j\|_2^2 > \frac{c}{\rho_j}$ for samples where $\rho_j \geq \bar{\rho}$, and that $\|\Delta_j\|_2^2 < \frac{c}{\rho_j}$ for samples where $\rho_j < \bar{\rho}$, for some $c > 0$. Then the BC-IR estimator has lower variance than the IS estimator:* $\mathbb{V}(X_{\mathrm{BC}} \mid B) < \mathbb{V}(X_{\mathrm{IS}} \mid B)$.

Theorem 3.4. *Assume ρ and the magnitude of the update $\|\Delta\|_2^2$ are independent*

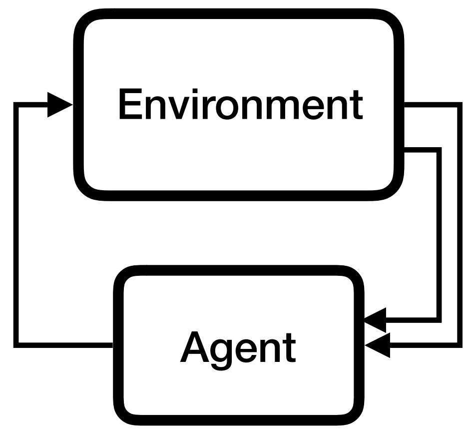
$$\mathbb{E}[\rho_j \|\Delta_j\|_2^2 \mid B] = \mathbb{E}[\rho_j \mid B]\, \mathbb{E}[\|\Delta_j\|_2^2 \mid B]$$

Then the BC-IR estimator will have equal or lower variance than the IS estimator.

Future Directions

Future Directions - Sampling Policy

Interact with Environment:



Add $\{\bar{\rho}_t, s_t, a_t, s_{t+1}\}$ **to** B

$$\bar{\rho}_t = \frac{d(s_t, a_i)}{b(s_t \mid a_t)}$$

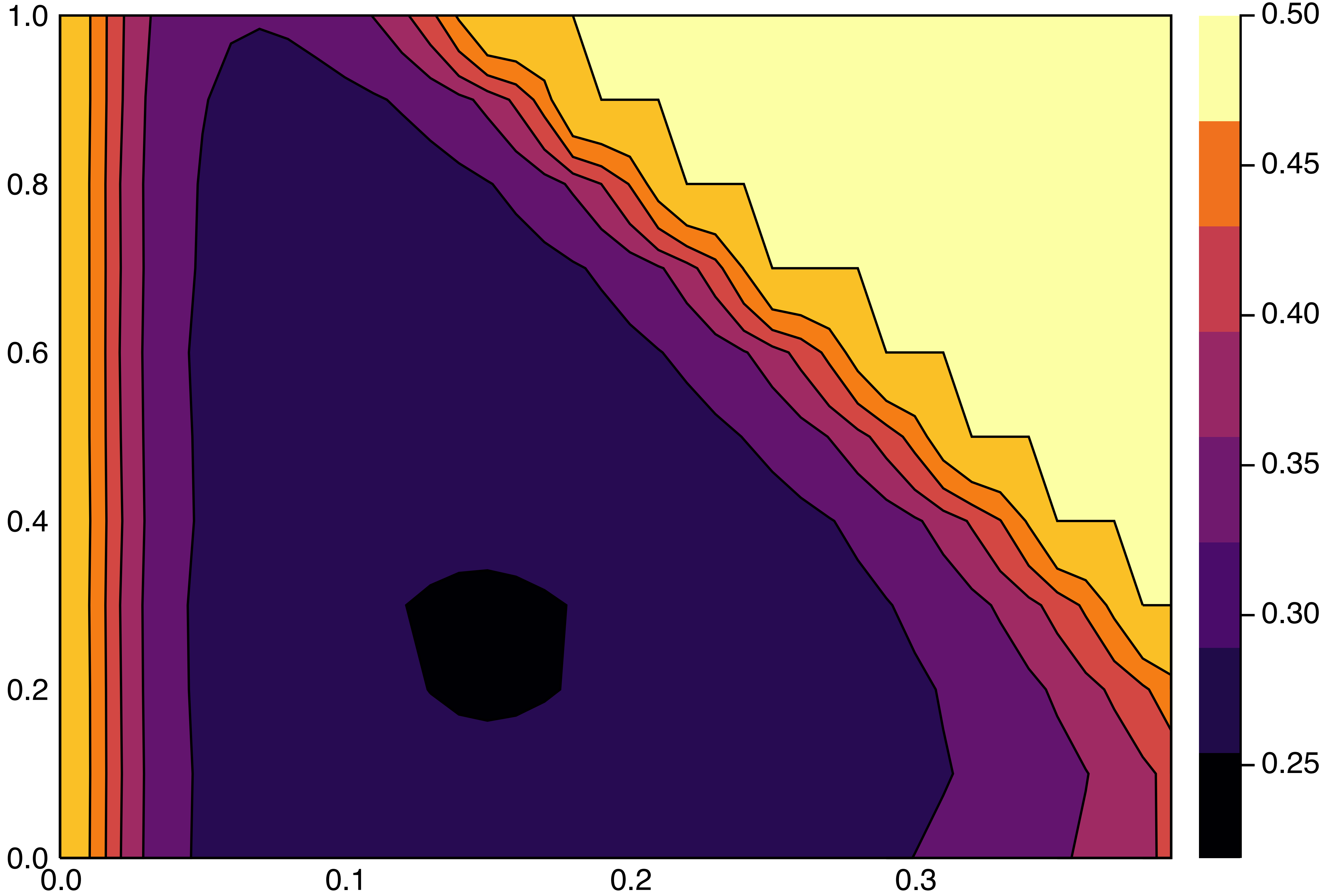
Sample Minibatch:

$\{\pi_{sample}(a_i \mid s_i), s_i, a_i, s_i'\}$ **with** $\mathbb{P}\left\{ \frac{\bar{\rho}_i}{\sum_{j=1}^{|B|} \bar{\rho}_j} \mid B \right\}$

Update Parameters:

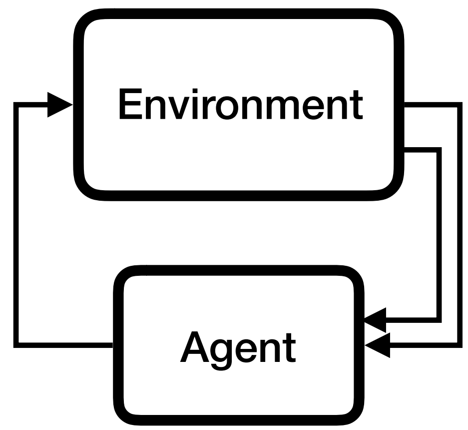
$$\Delta\theta = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i \mid s_i)}{d(s_i, a_i)} \delta_i \nabla_\theta V(s_i; \theta)$$

Future Directions - Sampling Policy



Future Directions - Multi-step learning

Interact with Environment:



Add $\{\bar{\rho}_t, s_t, a_t, s_{t+1}\}$ **to** B

$$\bar{\rho}_t = \rho_{t:t+k} = \prod_{i=t}^{t+k} \rho(a_i \mid s_i)$$

Sample Minibatch:

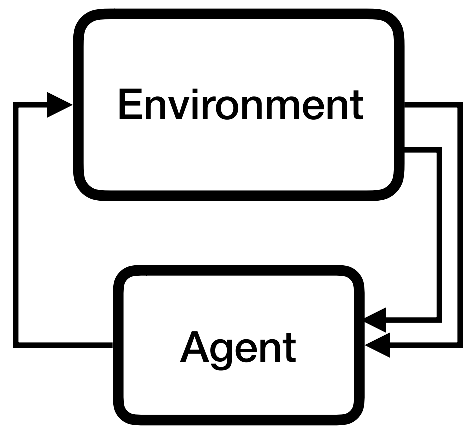
$$\{\bar{\rho}_i, s_i, a_i, s_i'\} \text{ with } \mathbb{P}\left\{ \frac{\bar{\rho}_i}{\sum_{j=1}^{|B|} \bar{\rho}_j} \mid B \right\}$$

Update Parameters:

$$\Delta\theta = \frac{1}{n} \sum_{i=1}^{n} \delta_i^k \nabla_\theta V(s_i; \theta)$$

Future Directions - state distributions

Interact with Environment:



Add $\{\bar{\rho}_t, s_t, a_t, s_{t+1}\}$ **to** B

$$\bar{\rho}_t = \frac{\pi(s_t \,|\, a_i)\hat{\mu}_\pi(s_t)}{b(s_t \,|\, a_t)\hat{\mu}_b(s_t)}$$

Sample Minibatch:

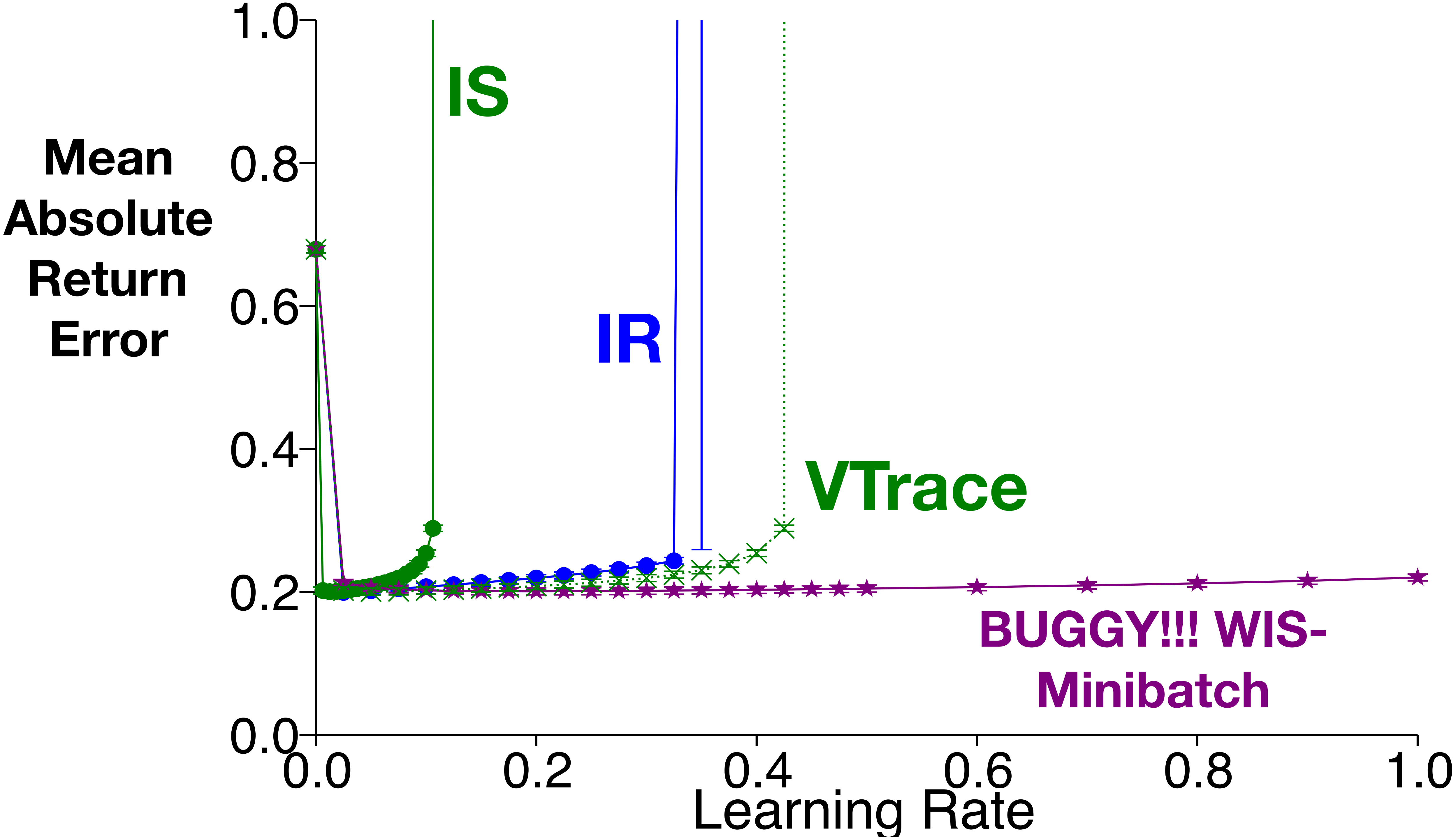
$\{\bar{\rho}_i, s_i, a_i, s_i'\}$ **with** $\mathbb{P}\left\{\frac{\bar{\rho}_i}{\sum_{j=1}^{|B|}\bar{\rho}_j} \,|\, B\right\}$

Update Parameters:

$$\Delta\theta = \frac{1}{n}\sum_{i=1}^{n}\delta_i \nabla_\theta V(s_i;\theta)$$

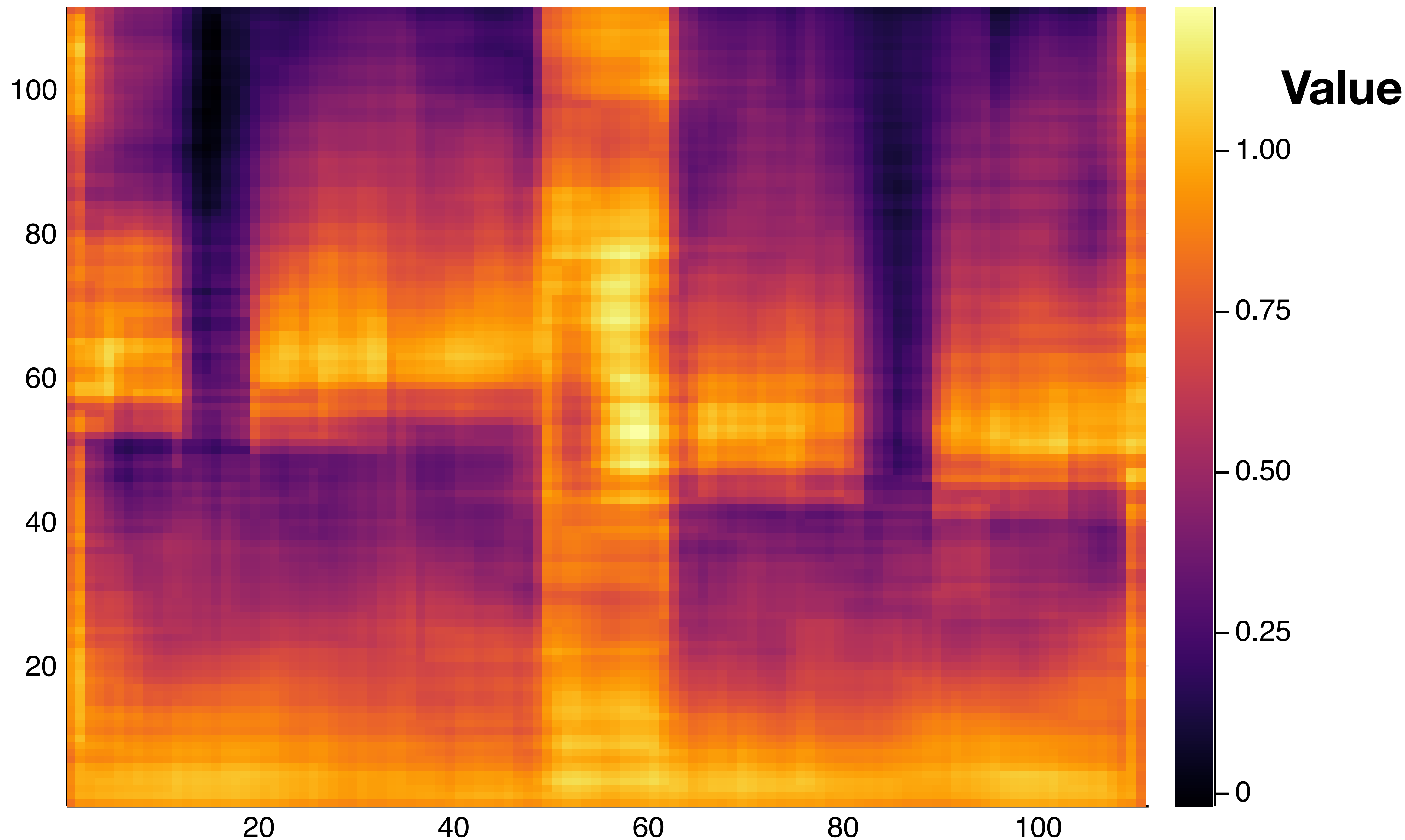
Extra Results

Four Rooms Cont LR sens



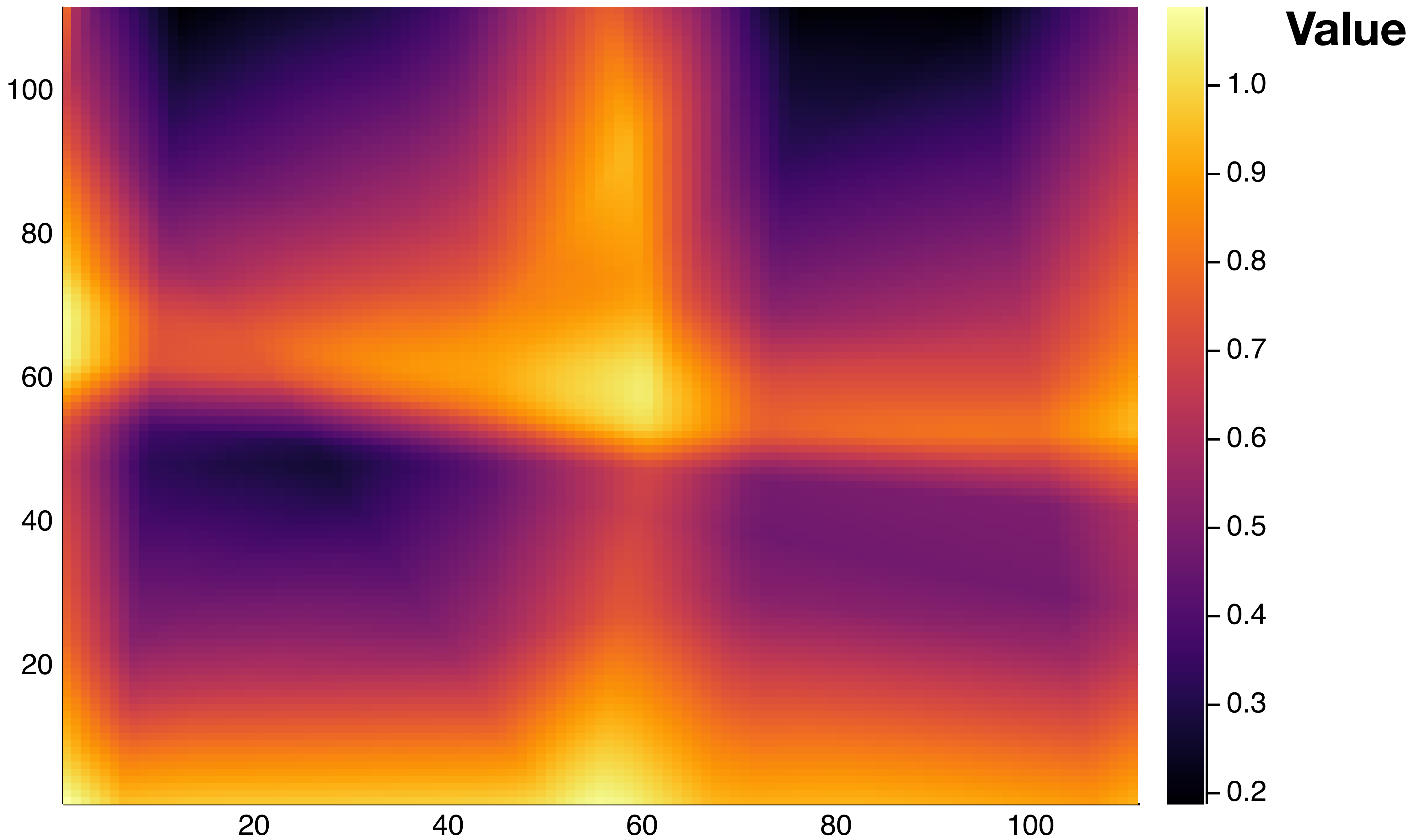
Learned Value Function CFR

IS (Tile Coded)

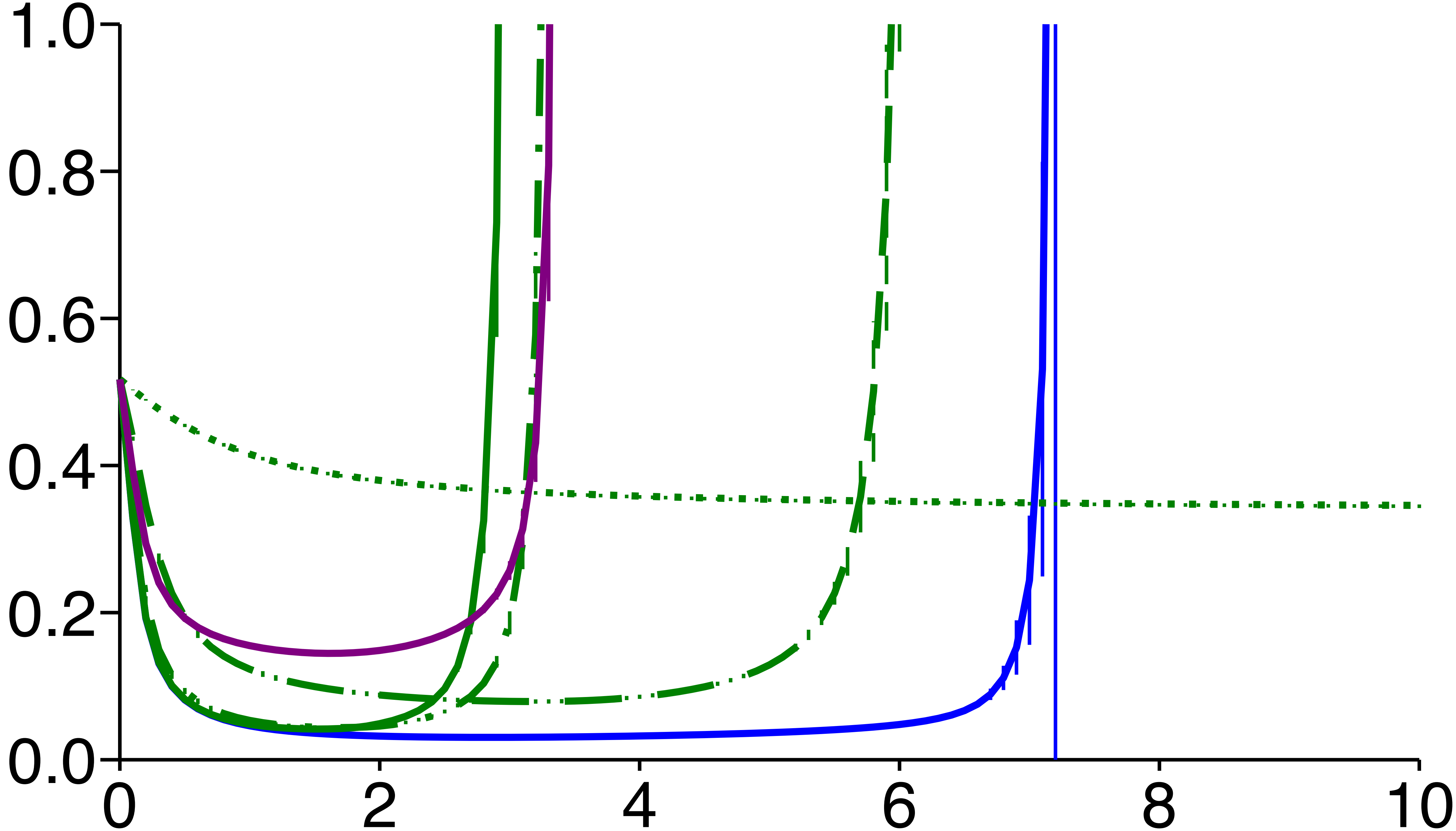


Learned Value Function CFR

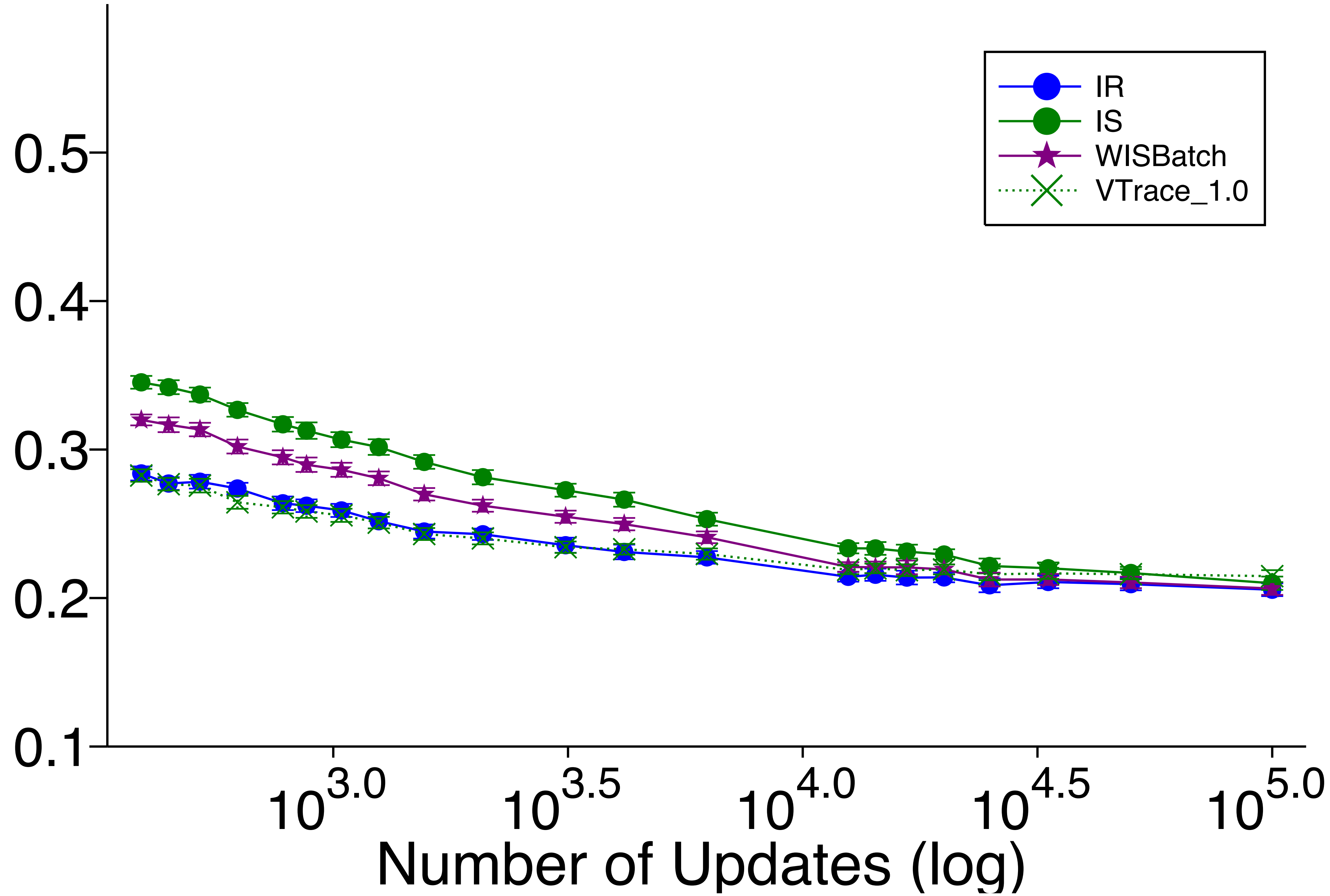
IS (Artificial Neural Network)



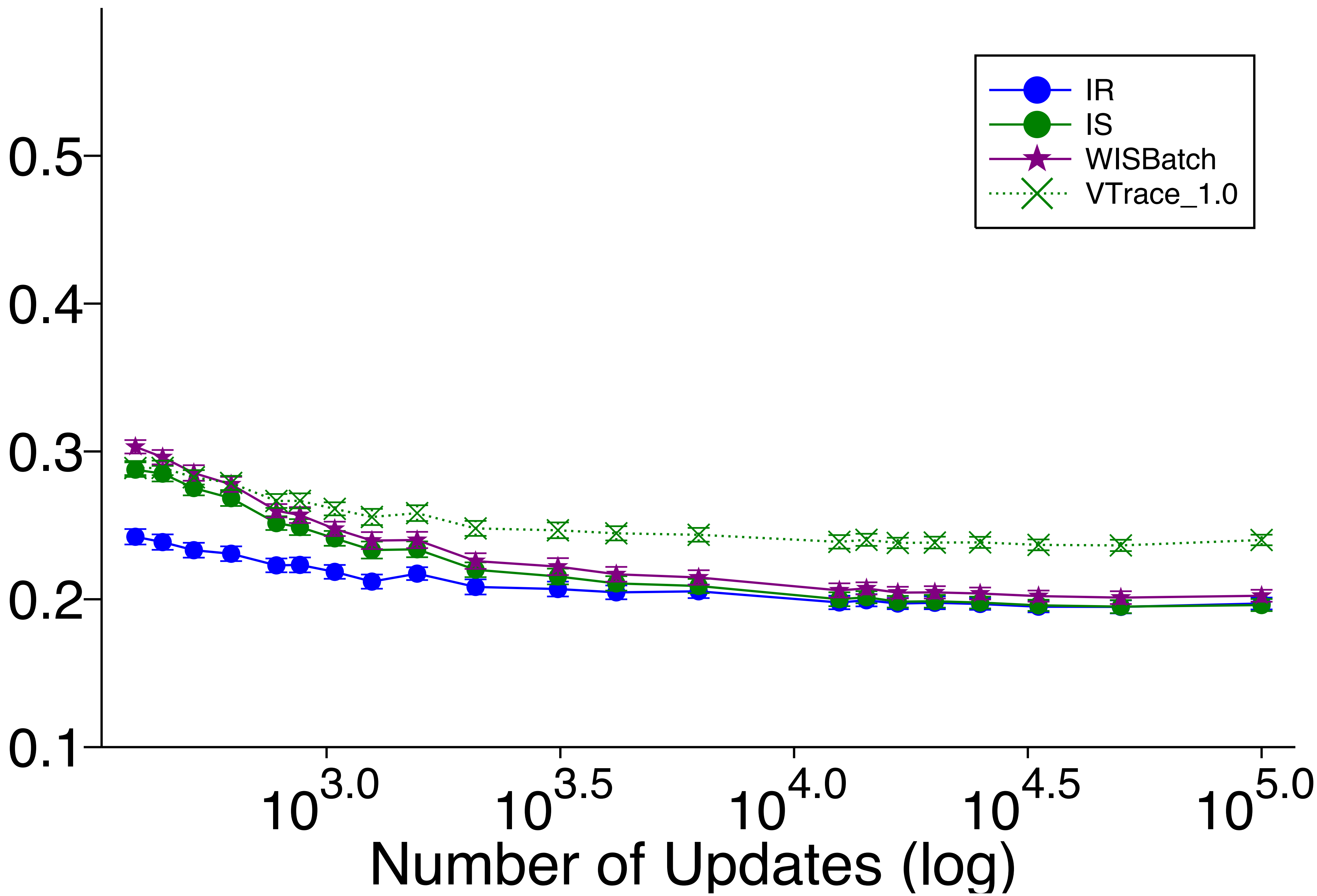
Future Directions - Odd Behavior of Bias



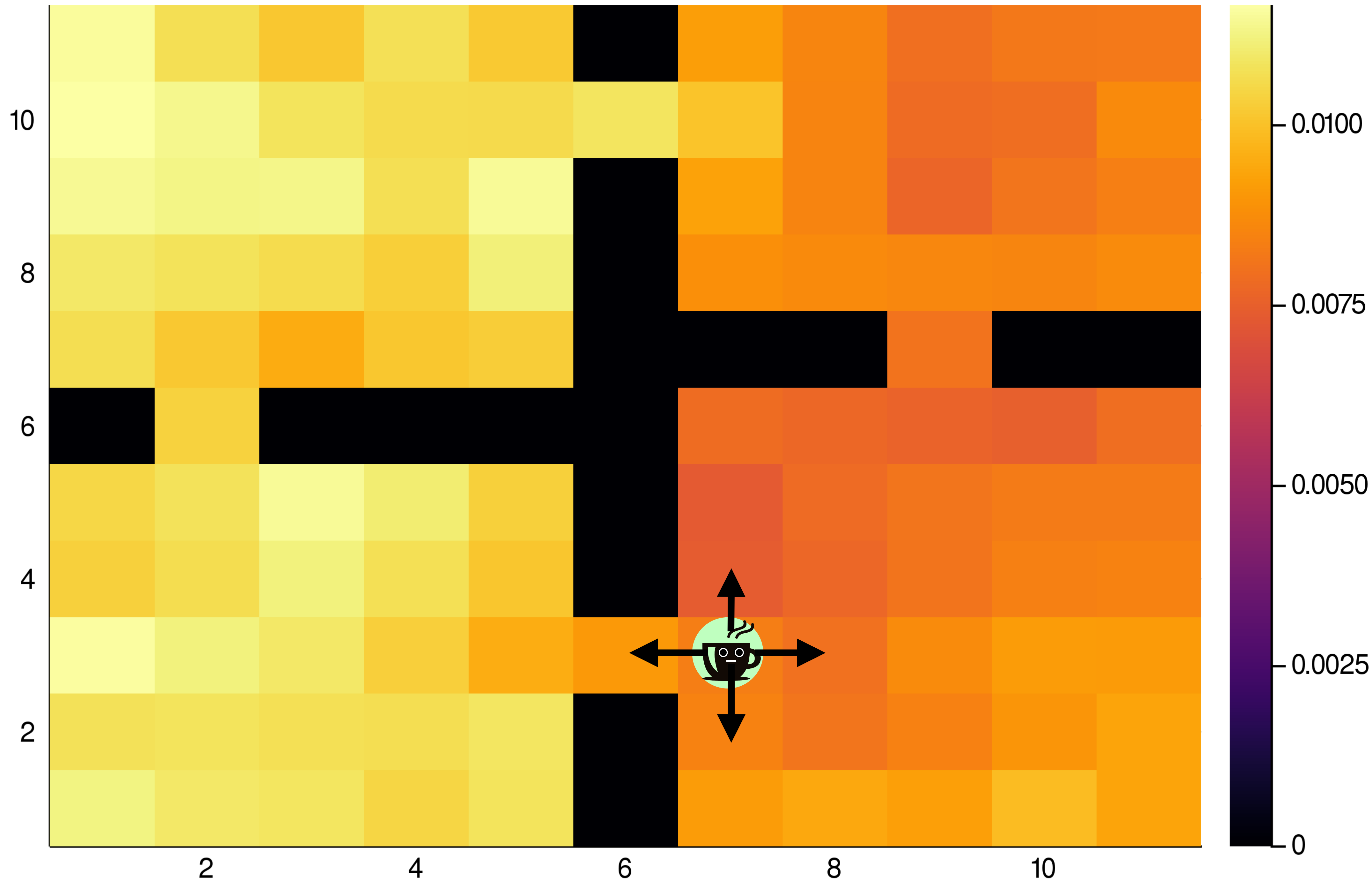
Future Directions - Odd Behavior of Bias



Odd Behavior of Bias (VTrace)



Four Rooms



Four Rooms

Evaluation:

- Estimate value function using dynamic programming

Behavior:

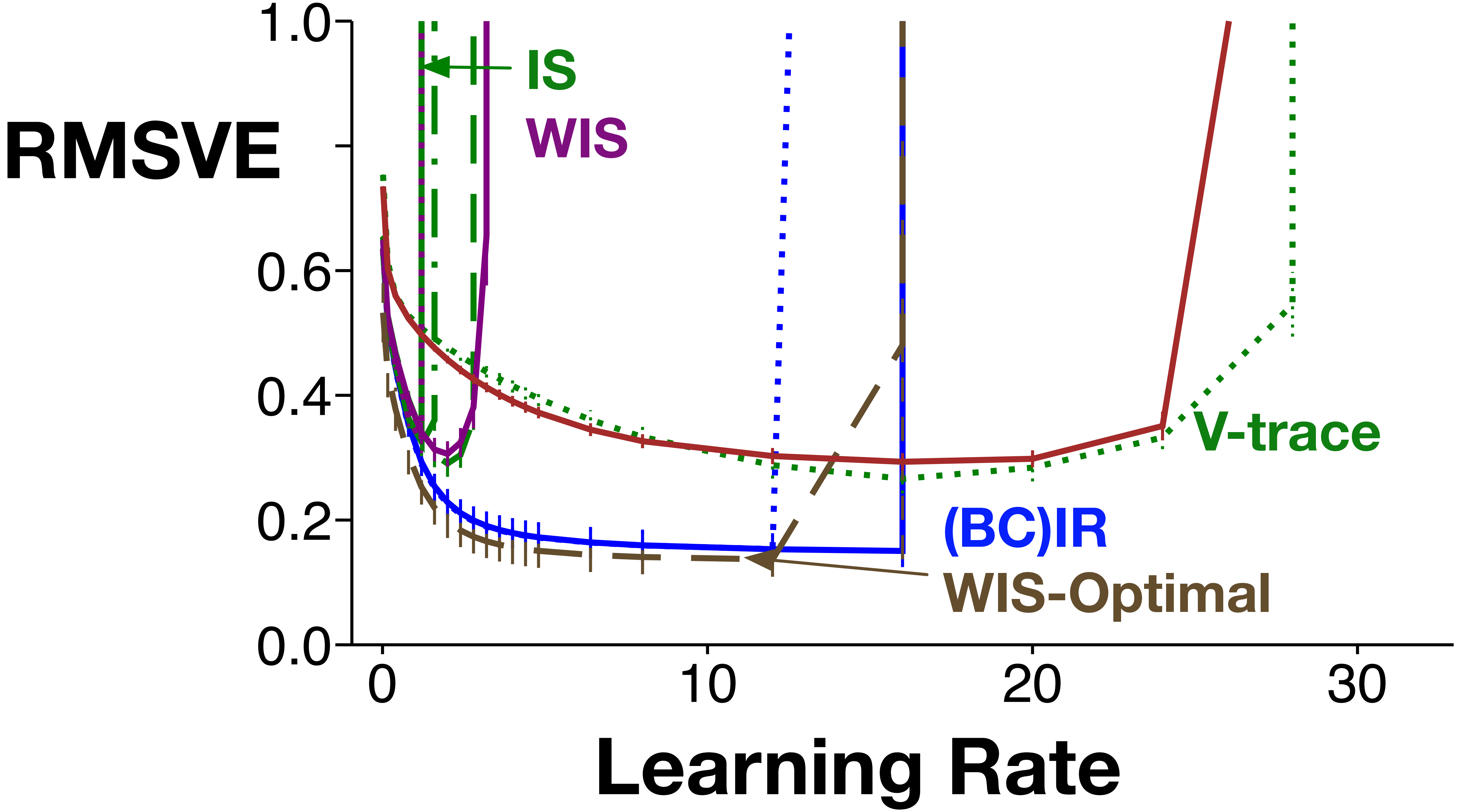
$$b(\cdot \,|\, s \notin S_{hv}) = 0.25$$

$$b(a \,|\, s \in S_{hv}) = \begin{cases} 0.1 & \textbf{if } a = down \\ \frac{0.9}{3} & \textbf{o.w.} \end{cases}$$

Target:

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

Four Rooms - 31,250 updates over 500k interactions



Four Rooms - 500k updates