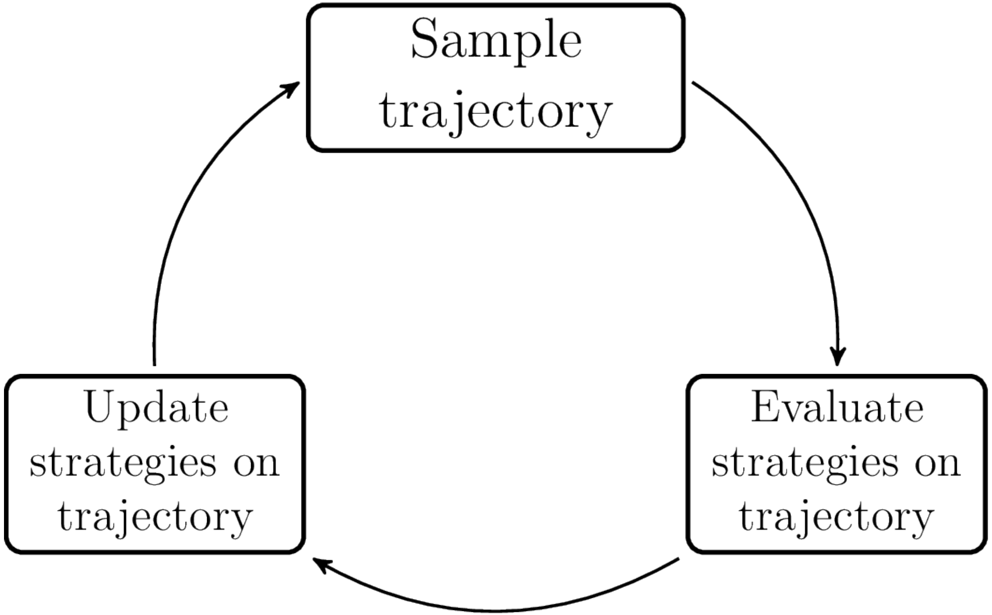
[1] DeepMind
[2] UNIVERSITY OF ALBERTA

Low-Variance and Zero-Variance Baselines in Extensive-Form Games

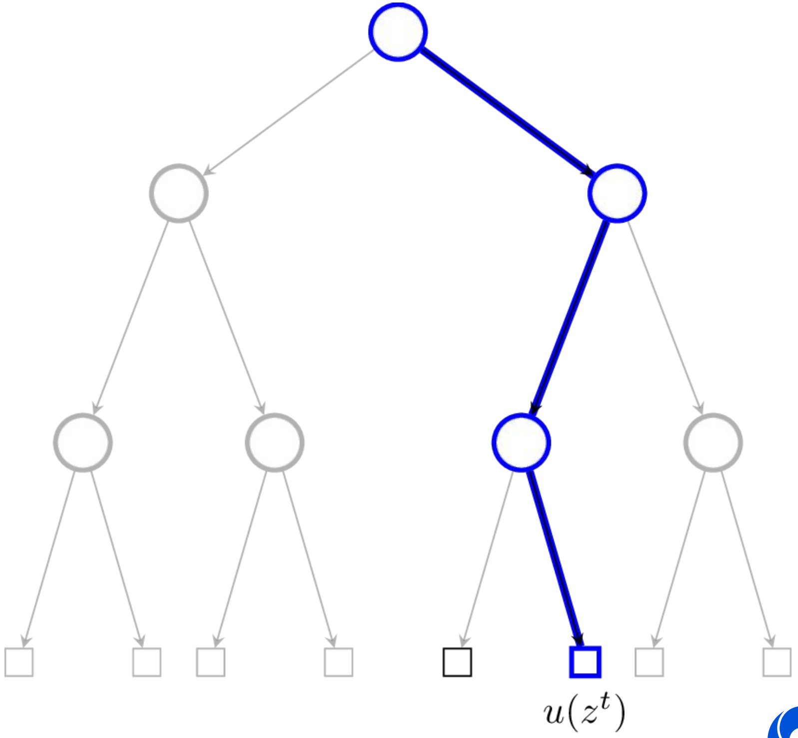
Trevor Davis[2,*], Martin Schmid[1], Michael Bowling[1,2]

*Work done during internship at DeepMind

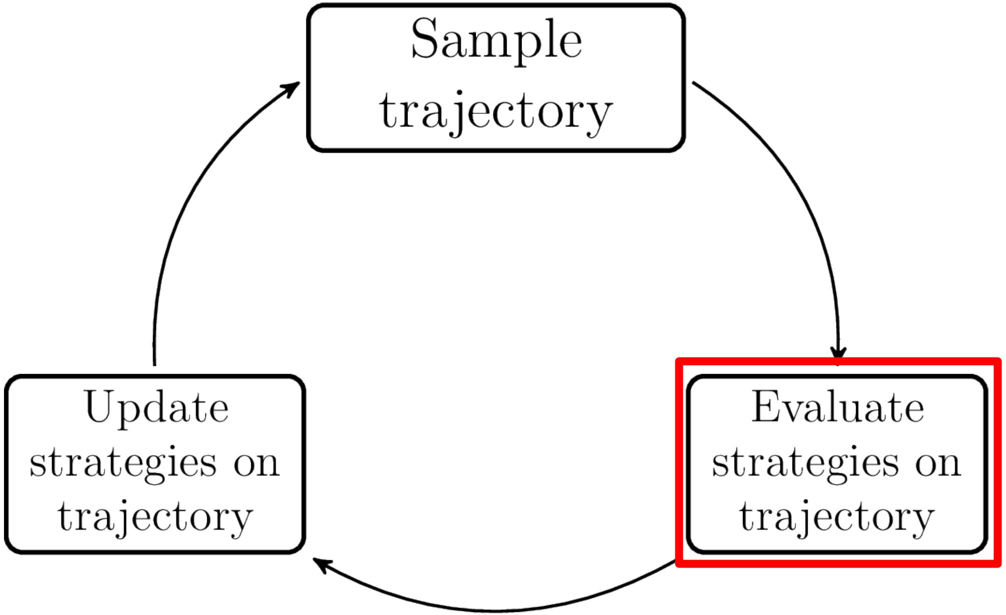
Monte Carlo game solving



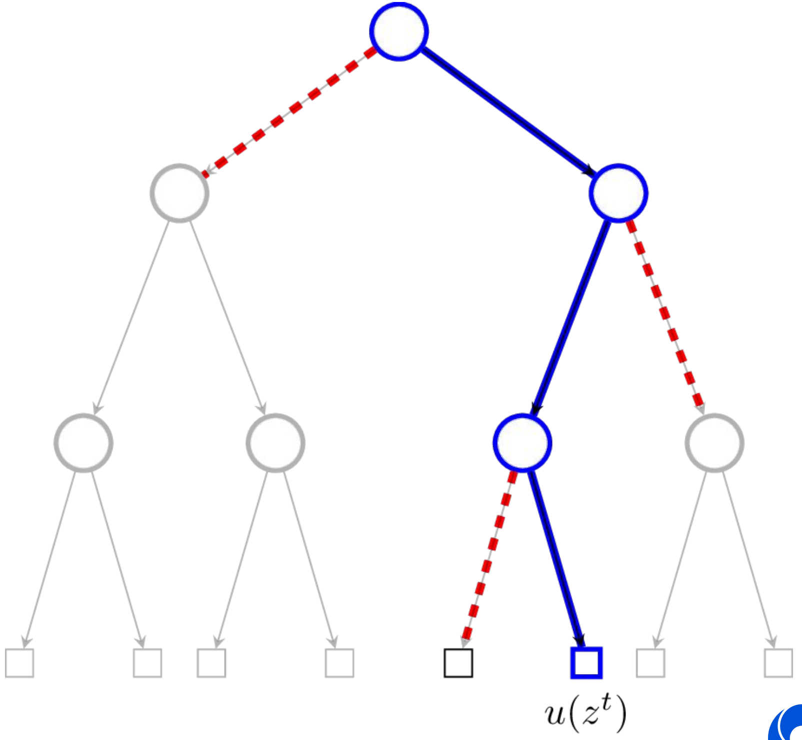
Extensive-form games (EFGs)



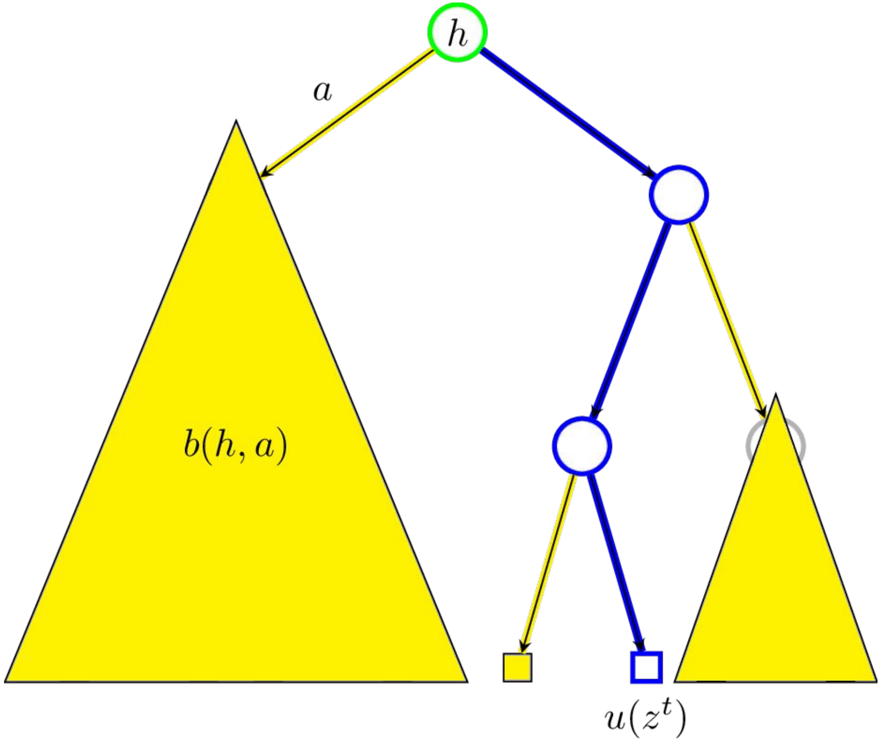
Monte Carlo game solving



Extensive-form games (EFGs)

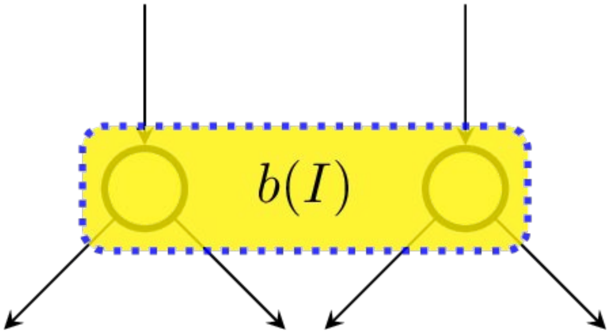


Baseline functions - evaluating unsampled actions

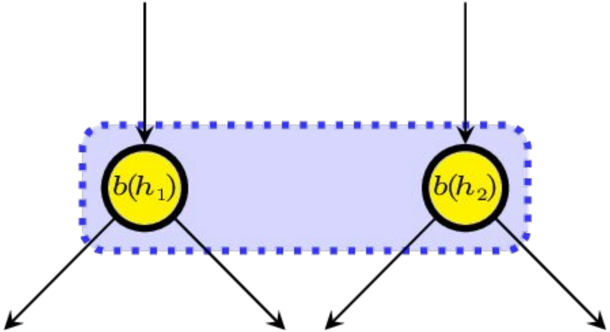


Our Contribution

VR-MCCFR
(Schmid et al., AAAI 2019)

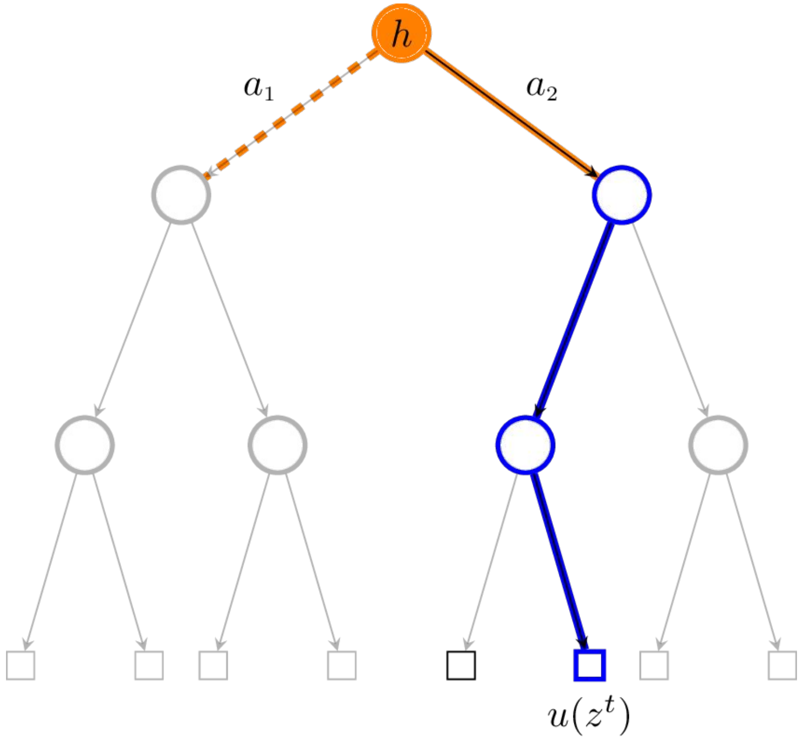


This work



- Lower variance, faster convergence
- Provable zero-variance samples

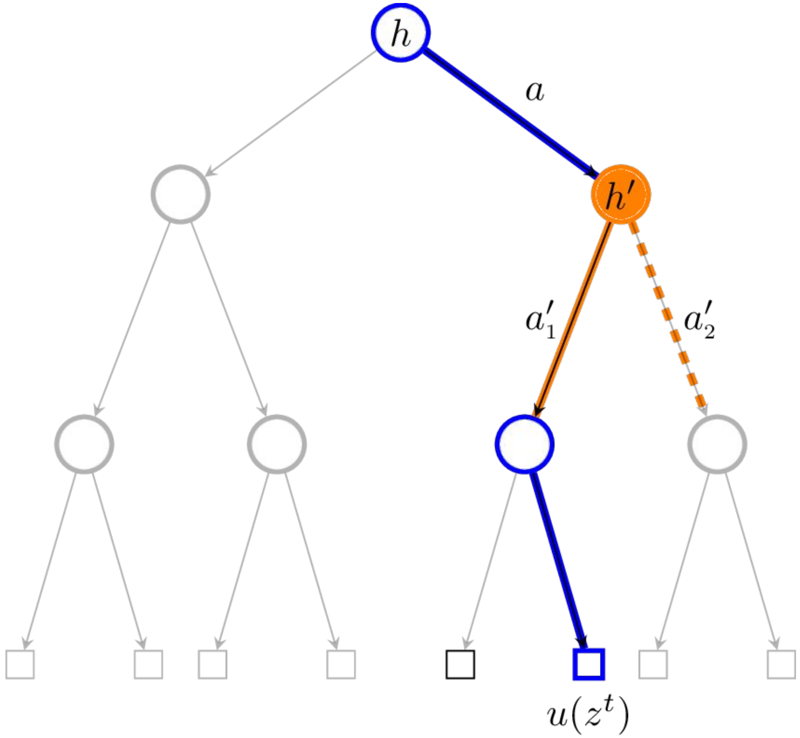
Monte carlo evaluation



Unbiased updates at *h*

$$\hat{u}(h,a) = \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]} u(z^t)$$

Monte Carlo evaluation



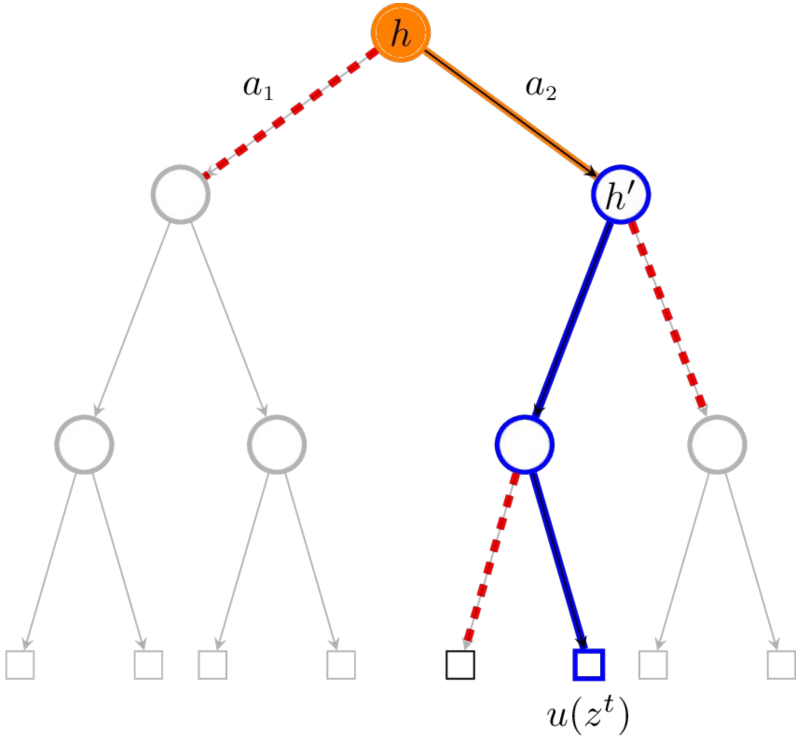
Unbiased updates at *h*

$$\hat{u}(h,a) = \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]} u(z^t)$$

$$= \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]} \hat{u}_b(h')$$

where $\hat{u}_b(h') := \sum_{a'} \Pr[h' \to a'] \hat{u}_b(h', a')$

Monte Carlo evaluation



Unbiased updates at *h*

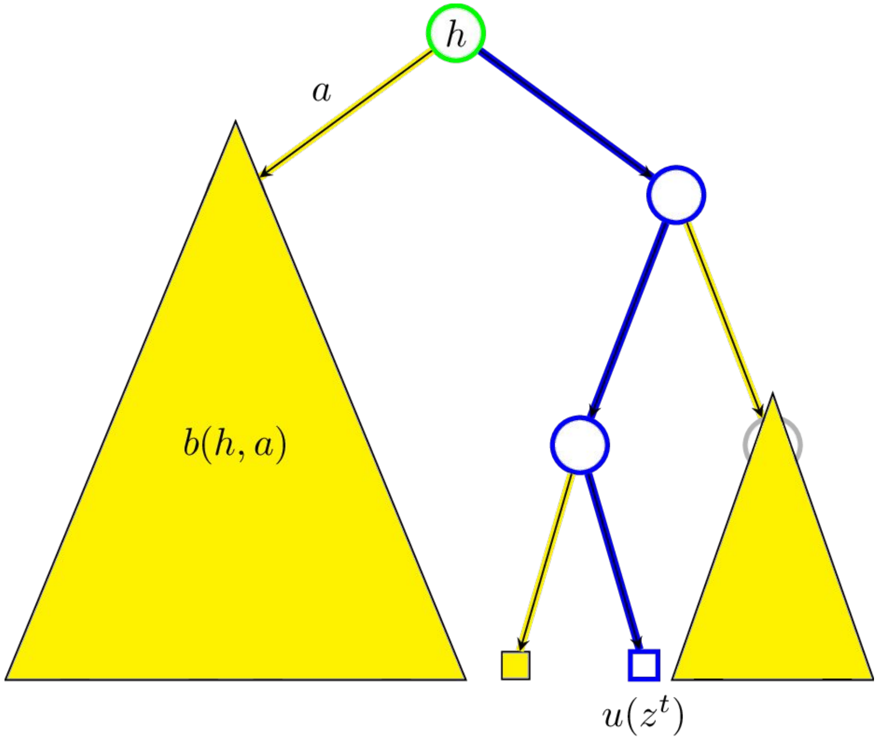
$$\hat{u}(h,a) = \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]} u(z^t)$$

$$= \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]} \hat{u}_b(h')$$

where $\hat{u}_b(h') := \sum_{a'} \Pr[h' \to a'] \hat{u}_b(h', a')$

Unsampled actions: $\hat{u}(h,a) = 0$

Baseline functions



$$b(h, a) \approx \mathbb{E}[u(h, a)]$$

Evaluation with baseline

Without baseline:

$$\hat{u}(h, a) = \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]} \hat{u}(h')$$

$$\hat{u}(h') = \sum_{a'} \Pr[h' \to a'] \hat{u}(h', a')$$

Evaluation with baseline

Without baseline:

$$\hat{u}(h, a) = \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]}\hat{u}(h')$$

$$\hat{u}(h') = \sum_{a'} \Pr[h' \to a']\hat{u}(h', a')$$

Baseline correction:

$$\hat{u}_b(h, a) = \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]}\hat{u}_b(h') + \left(1 - \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]}\right) b(h, a)$$

Evaluation with baseline

Without baseline:

$$\hat{u}(h,a) = \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]}\hat{u}(h') \qquad \hat{u}(h') = \sum_{a'} \Pr[h' \to a']\hat{u}(h',a')$$

Baseline correction:

$$\hat{u}_b(h,a) = \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]}\hat{u}_b(h') + \underbrace{\left(1 - \frac{\mathbb{1}(h \to a)}{\Pr[h \to a]}\right) b(h,a)}_{\mathbb{E}[\dots] = 0}$$

(control variate)

Theoretical results

Theorem 1: baseline-corrected values are unbiased:

$$\mathbb{E}[\hat{u}_b(h,a)] = \mathbb{E}[u(h,a)]$$

Theorem 2: each baseline-corrected value $\hat{u}_b(h,a)$ has variance bounded by a sum of squared prediction errors in the subtree rooted at a

$$\mathrm{Var}[\hat{u}_b(h,a)] \leq \frac{1}{\Pr[h \to a]} \sum_{\substack{h',a' \in \\ \mathrm{subtree}(h,a)}} \Pr[h,a \to h',a'](b(h',a') - \mathbb{E}[u(h',a')])^2$$

Baseline function selection

We want $b(h, a) \approx \mathbb{E}[u(h, a)]$

Learned history baseline:

We know $\mathbb{E}[\hat{u}_b(h, a)] = \mathbb{E}[u(h, a)]$

Set $b(h, a)$ to average of previous samples $\hat{u}_b(h, a)$

Baseline function selection

We want $b(h,a) \approx \mathbb{E}[u(h,a)]$

Learned history baseline:

We know $\mathbb{E}[\hat{u}_b(h,a)] = \mathbb{E}[u(h,a)]$

Set $b(h,a)$ to average of previous samples $\hat{u}_b(h,a)$

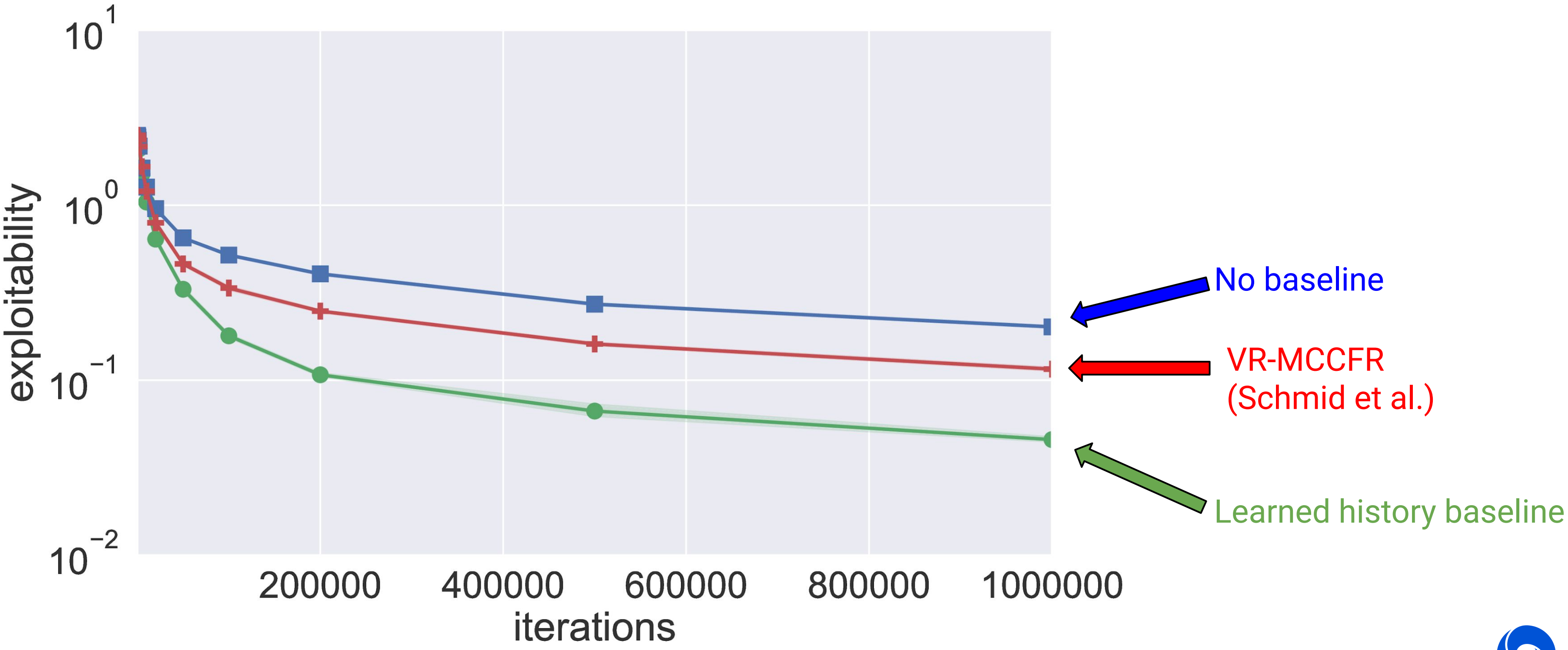
Note: $\mathbb{E}[u(h,a)]$ depends on strategies – not stationary

$\therefore b(h,a)$ is *not* an unbiased estimate of current expectation

$\hat{u}_b(h,a)$ still unbiased

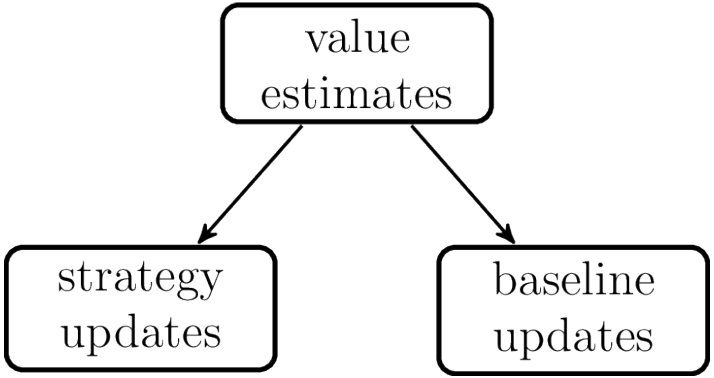
Baseline convergence evaluation

Leduc poker, Monte Carlo Counterfactual Regret Minimization (MCCFR+)



Predictive baseline

Updating with learned history baseline:

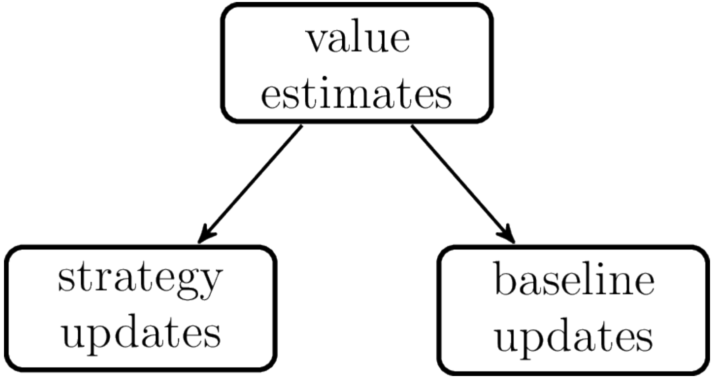


Optimal baseline depends on strategy update:

$$b(h, a) = \mathbb{E}[u(h, a)]$$
$$= \sum_{a'} \Pr[h' \to a']\mathbb{E}[u(h', a')]$$

Predictive baseline

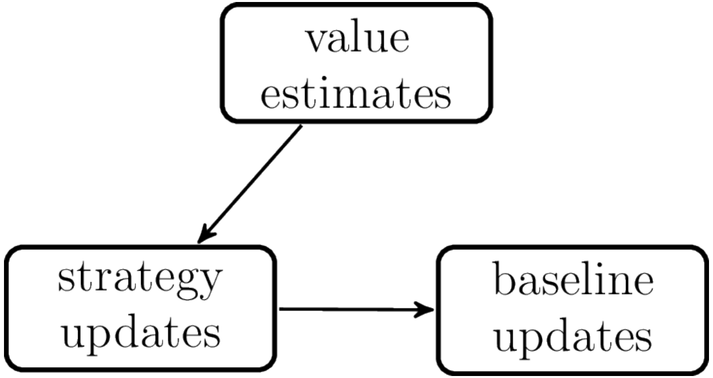
Updating with learned history baseline:



Optimal baseline depends on strategy update:

$$b(h,a) = \mathbb{E}[u(h,a)]$$
$$= \sum_{a'} \Pr[h' \to a']\mathbb{E}[u(h',a')]$$

Use strategy to update baseline:



Recursively set

$$b(h,a) = \sum_{a'} \Pr[h' \to a'] b(h',a')$$

Zero-variance updates

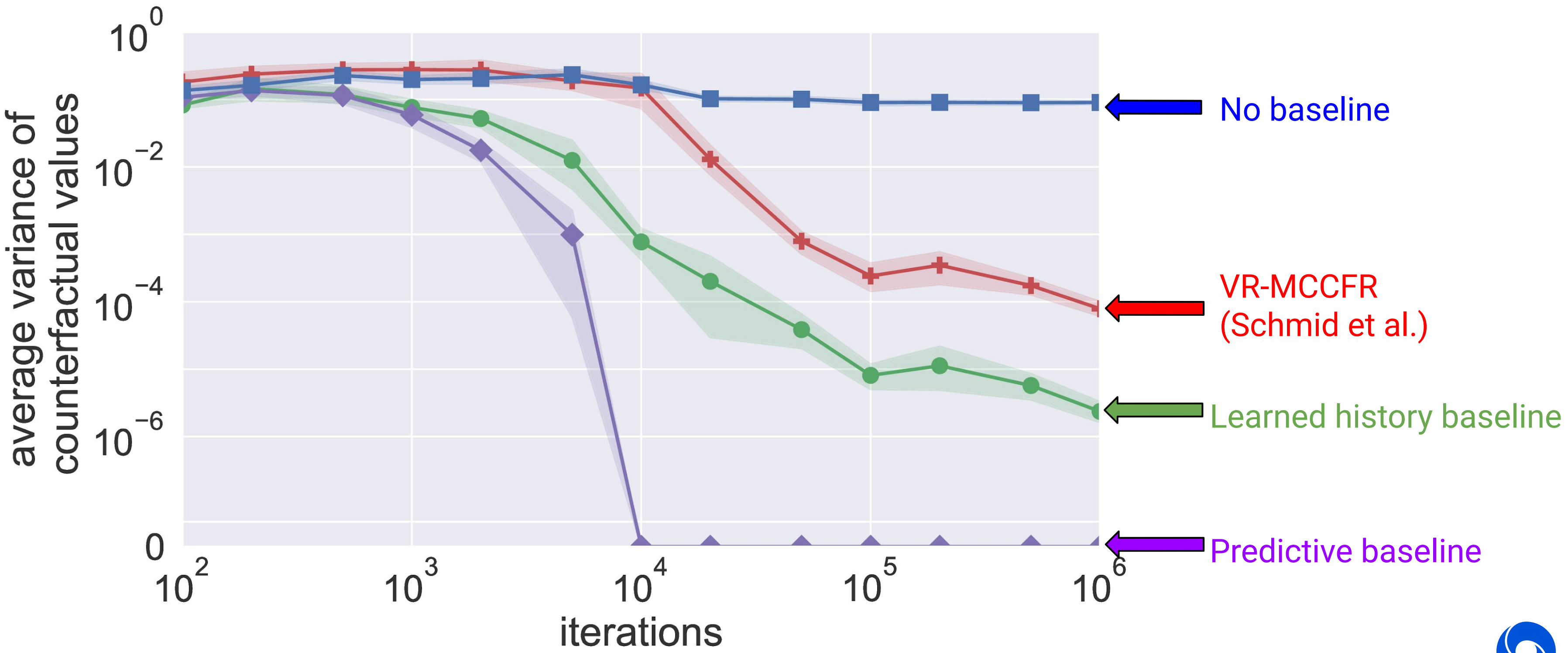
If:

- We use the predictive baseline
- We sample public outcomes
- All outcomes are sampled at least once

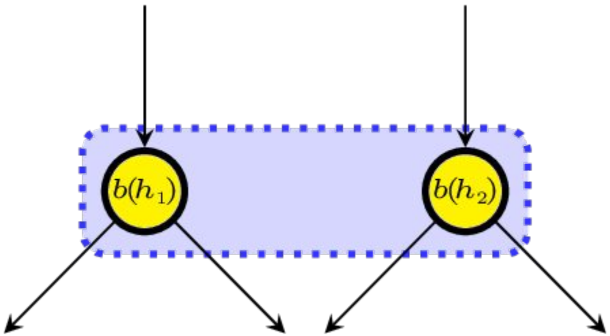
Theorem: the baseline-corrected values $\hat{u}_b(h, a)$ have zero variance

Baseline variance evaluation

Leduc poker, Monte Carlo Counterfactual Regret Minimization (MCCFR+)



Conclusion



- Lower variance, faster convergence
- Provable zero-variance samples