NEURAL INFORMATION PROCESSING SYSTEMS

PRINCETON UNIVERSITY

Hindsight Credit Assignment

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Nicolas Heess, Mohammad G. Azar, Bilal Piot, Hado van Hasselt, Satinder Singh, Greg Wayne, Doina Precup, Rémi Munos
DeepMind
`{harutyunyan, wdabney, munos}@google.com`

"Tony" Runzhe Yang

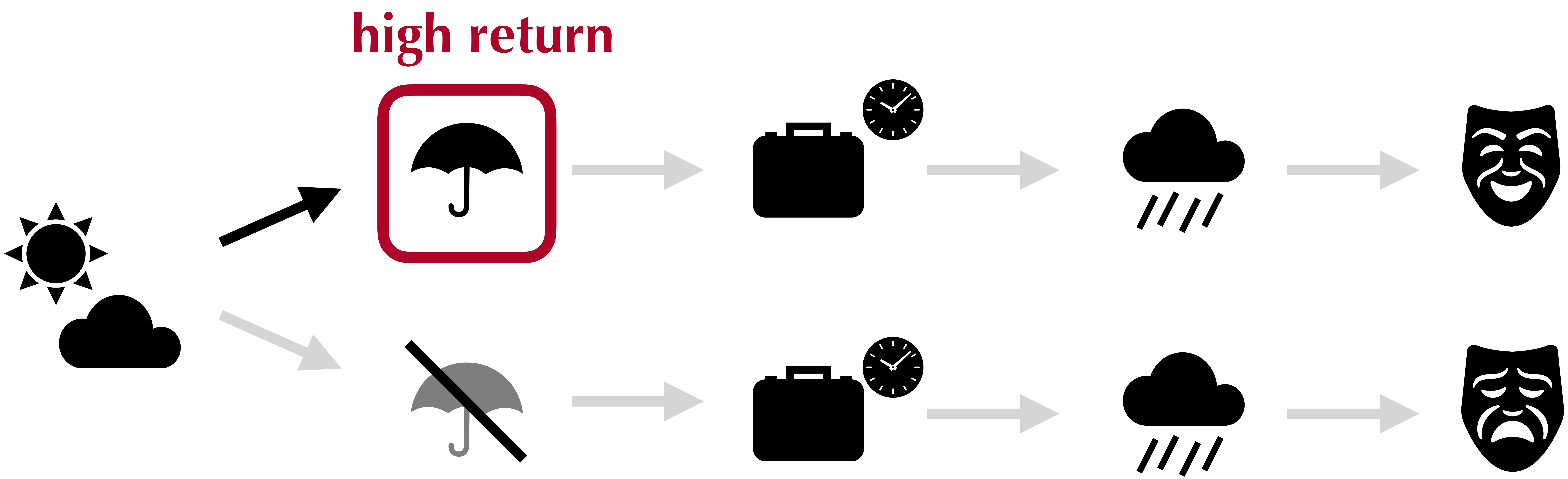
My 20th, 2020

https://runzhe-yang.science

Value Function Problem

$$V^\pi(x) \stackrel{\text{def}}{=} \mathbb{E}_{\tau \sim \mathcal{T}(x,\pi)}\Big[Z(\tau)\Big], \qquad Q^\pi(x,a) \stackrel{\text{def}}{=} \mathbb{E}_{\tau \sim \mathcal{T}(x,a,\pi)}\Big[Z(\tau)\Big].$$

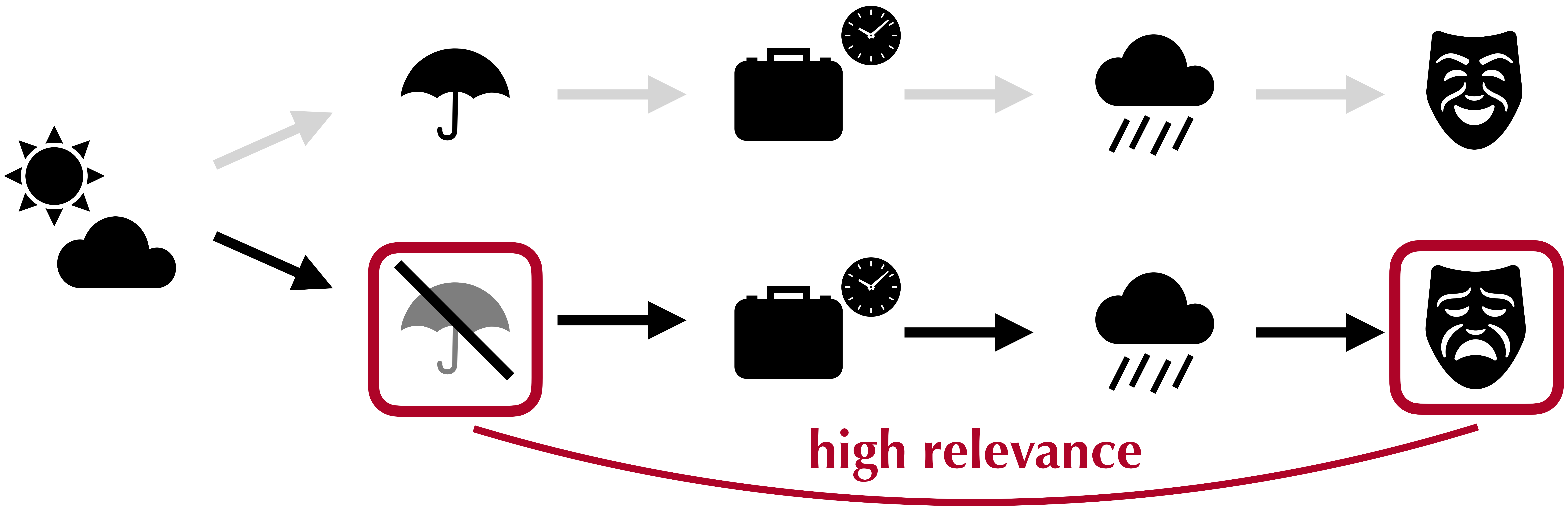
"how does the **current action** affect **future outcomes**?"



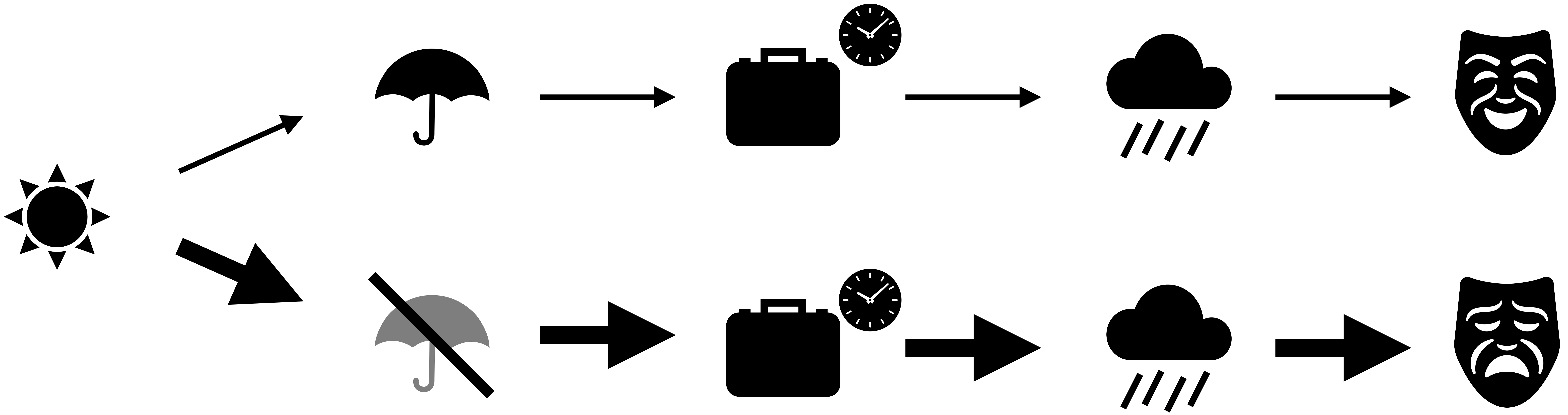
Credit Assignment Problem

$$I(A_t; f(\tau_{t:\infty})|X_t = x) = \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}\left[\log\left(\frac{\mathbb{P}(A = A_t|f(\tau) = f(\tau_{t:\infty}), X_t = x)}{\mathbb{P}(A = A_t|X_t = x)}\right)\right]$$

"given an **outcome**, how *relevant* were **past decisions**?"

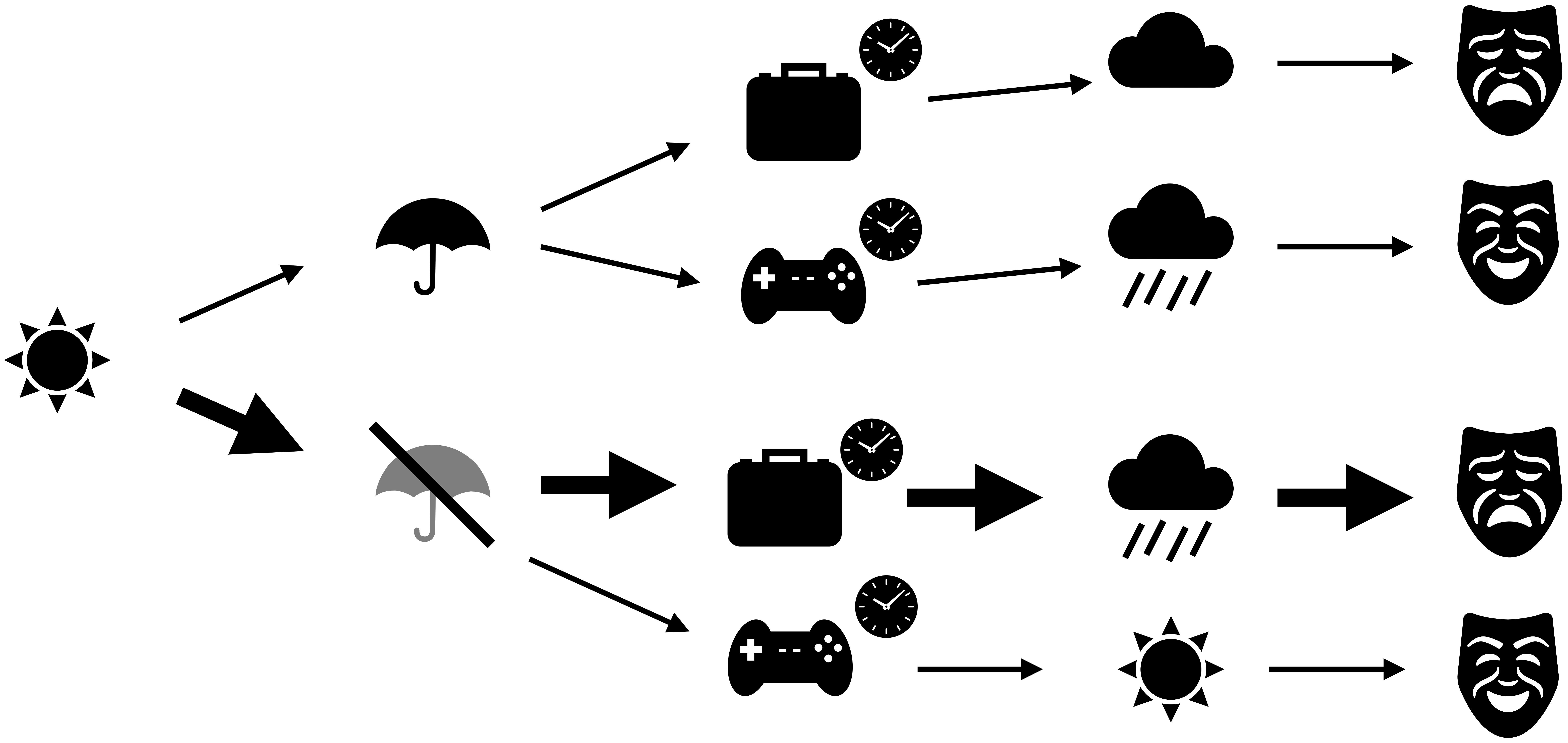


Credit Assignment Problem - Why is it important?



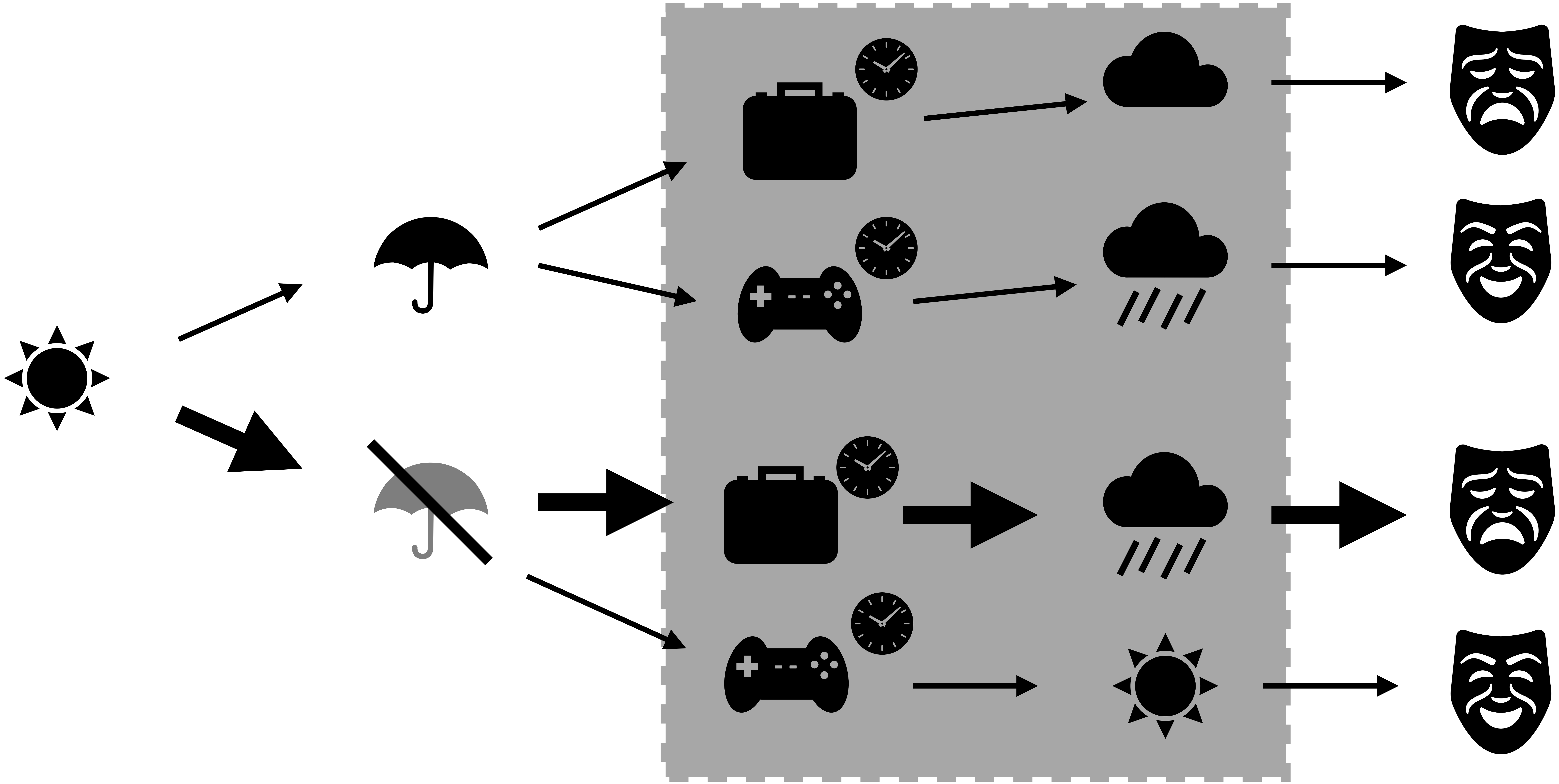
Rare events require an infeasible number of samples to obtain an accurate estimate.

Credit Assignment Problem - Why is it challenging?



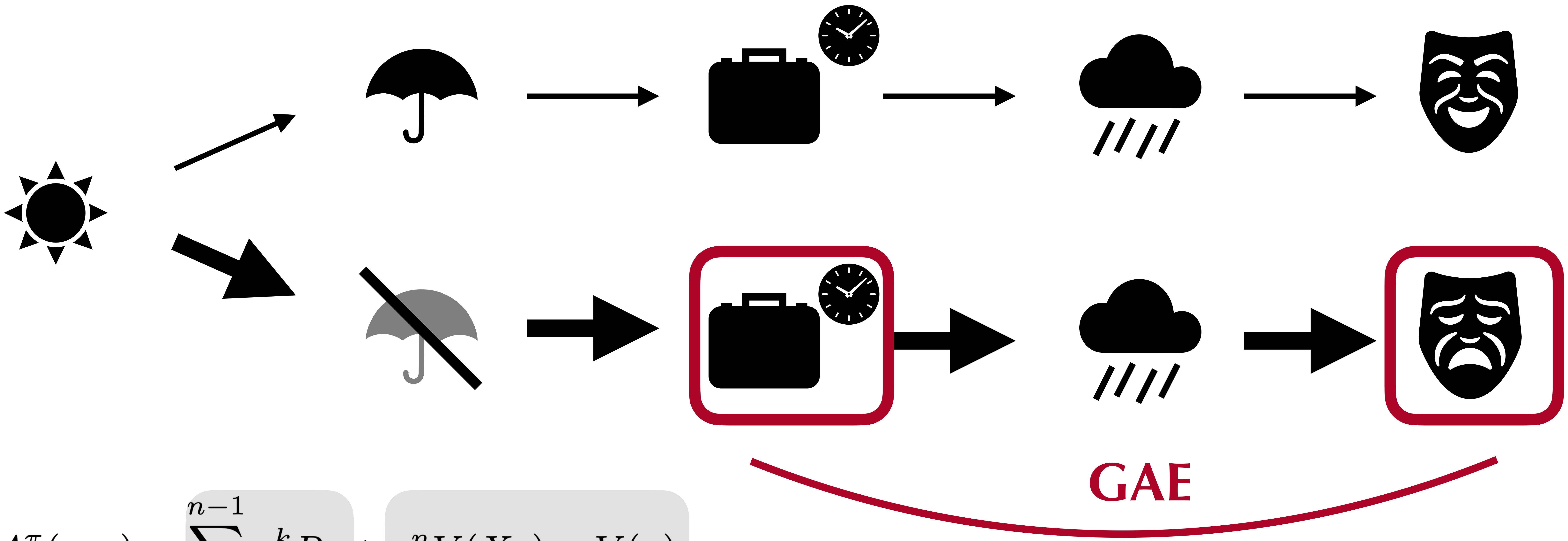
Issue 1: Variance - low sample efficiency

Credit Assignment Problem - Why is it challenging?



Issue 2: Partial observability - cannot bootstrap.

Credit Assignment Problem - Why is it challenging?



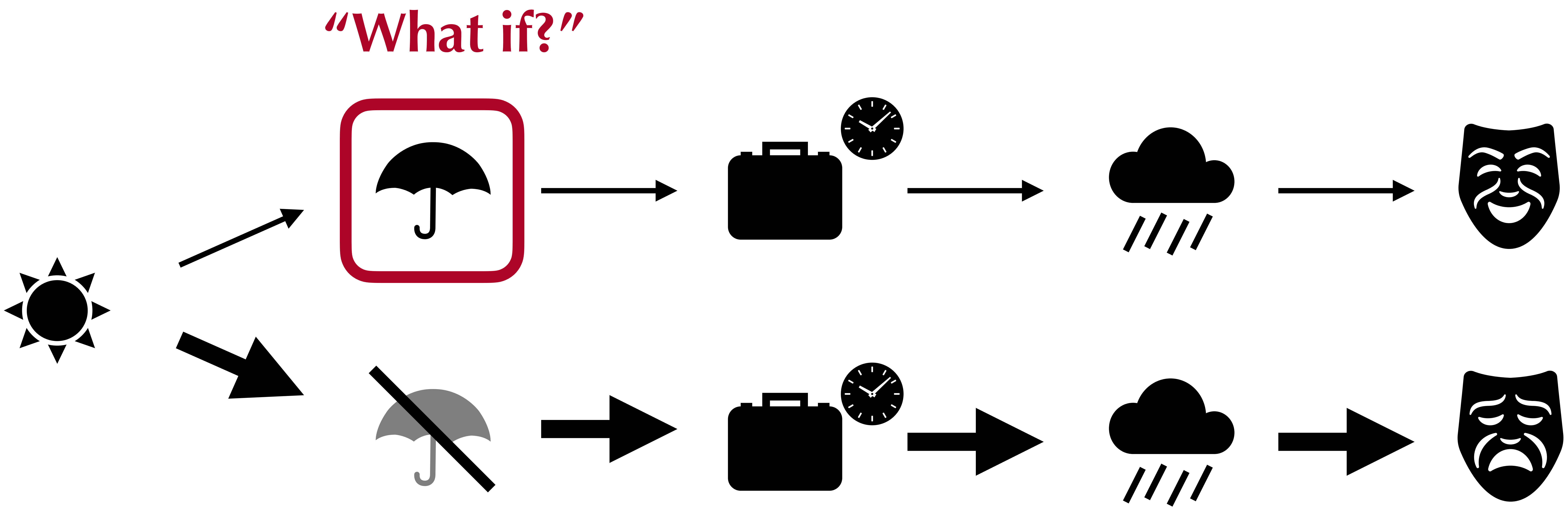
$$A^{\pi}(x,a) \approx \sum_{k=0}^{n-1} \gamma^k R_k + \gamma^n V(X_n) - V(x).$$

variance

bias —> best *n*?

Issue 3: Time as a proxy - rely on *time* as the sole metric.

Credit Assignment Problem - Why is it challenging?

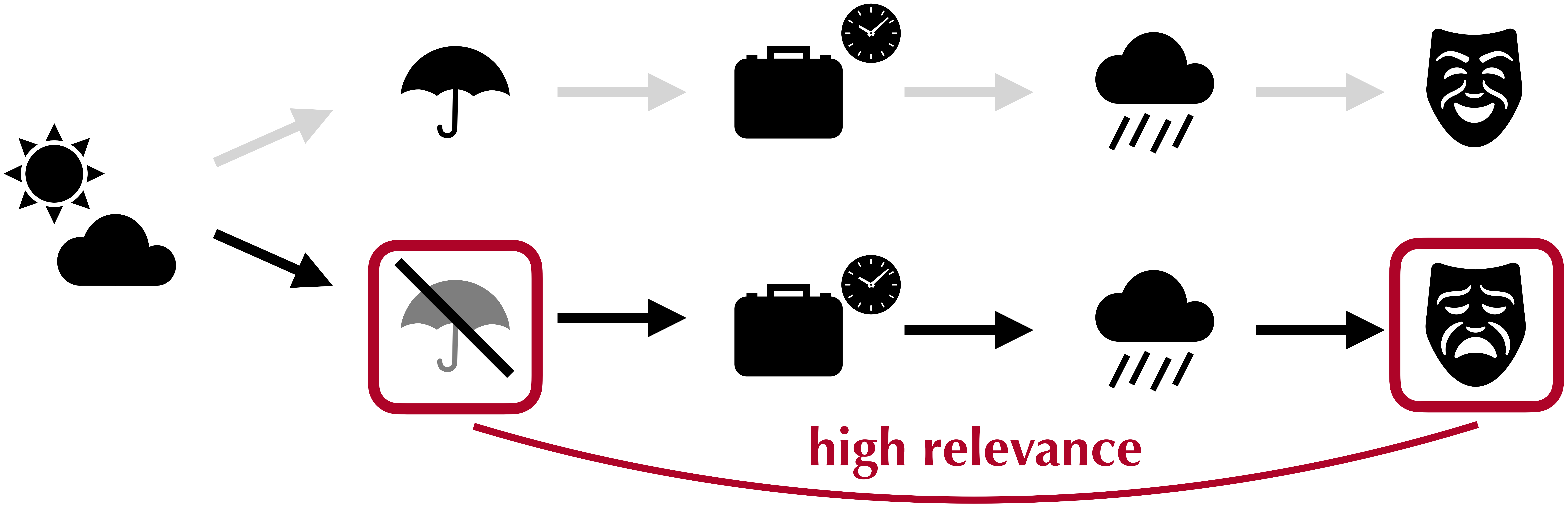


Issue 4: No counterfactuals - only update actions serendipitously occur.

Credit Assignment - Mutual Information Perspective

$$I(A_t; f(\tau_{t:\infty})|X_t = x) = \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}\left[\log\left(\frac{\mathbb{P}(A = A_t|f(\tau) = f(\tau_{t:\infty}), X_t = x)}{\mathbb{P}(A = A_t|X_t = x)}\right)\right]$$

"given an **outcome**, how *relevant* were **past decisions**?"



Credit Assignment - Mutual Information Perspective

$$I(A_t; f(\tau_{t:\infty}) | X_t = x) = \mathbb{E}_{\tau \sim \mathcal{T}(x,\pi)} \left[\log \left(\frac{\mathbb{P}(A = A_t | f(\tau) = f(\tau_{t:\infty}), X_t = x)}{\mathbb{P}(A = A_t | X_t = x)} \right) \right]$$

density ratio depicts relevance of *actions* and *outcomes* given states

$$\frac{h(a|x,\pi,f(\tau))}{\pi(a|x)}$$

Predictive Coding

can be learned by **InfoNCE** and other supervised learning method.

Future States

$$h_k(a|x,\pi,y) \stackrel{def}{=} \mathbb{P}_{\tau \sim \mathcal{T}(x,\pi)}(A_0 = a | X_k = y).$$

Future Returns

$$h_z(a|x,\pi,z) \stackrel{def}{=} \mathbb{P}_{\tau \sim \mathcal{T}(x,\pi)}(A_0 = a | Z(\tau) = z).$$

Credit Assignment - Conditioning on Future States

$$\frac{h(a|x,\pi,f(\tau))}{\pi(a|x)}$$

Predictive Coding

Future States

$$h_k(a|x,\pi,y) \stackrel{def}{=} \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|X_k = y).$$

Bayes' rule:

$$\frac{h_k(a|x,\pi,y)}{\pi(a|x)} = \frac{\mathbb{P}(X_k = y|X_0 = x, A_0 = a, \pi)}{\mathbb{P}(X_k = y|X_0 = x, \pi)} = \frac{\mathbb{P}_{\tau\sim\mathcal{T}(x,a,\pi)}(X_k = y)}{\mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(X_k = y)}.$$

any trajectory starts with *x*

> 1 when ***a*** and ***y*** are positively correlated

< 1 when ***a*** and ***y*** are negatively correlated

lower entropy

Credit Assignment - Conditioning on Future States

$$\frac{h(a|x,\pi,f(\tau))}{\pi(a|x)}$$

Predictive Coding

Future States

$$h_k(a|x,\pi,y) \stackrel{def}{=} \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|X_k = y).$$

Bayes' rule:

$$\frac{h_k(a|x,\pi,y)}{\pi(a|x)} = \frac{\mathbb{P}(X_k = y|X_0 = x, A_0 = a, \pi)}{\mathbb{P}(X_k = y|X_0 = x, \pi)} = \frac{\mathbb{P}_{\tau\sim\mathcal{T}(x,a,\pi)}(X_k = y)}{\mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(X_k = y)}.$$

any trajectory starts with *x*

Thm. 1

$$\Rightarrow \quad Q^\pi(x,a) = r(x,a) + \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}\left[\sum_{k\geq 1}\gamma^k \frac{h_k(a|x,X_k)}{\pi(a|x)} R_k\right].$$

counterfactual importance sampling

Credit Assignment - Conditioning on Future States

$$Q^\pi(x,a) = r(x,a) + \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}\Big[\sum_{k\geq 1}\gamma^k \frac{h_k(a|x,X_k)}{\pi(a|x)} R_k\Big].$$

counterfactual importance sampling

$$\Rightarrow \quad A^\pi(x,a) = r(x,a) - r^\pi(x) + \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}\Big[\sum_{k\geq 1}\Big(\frac{h_k(a|x,X_k)}{\pi(a|x)} - 1\Big)\gamma^k R_k\Big]$$

= 0, when irrelevant

Algorithm:

$$\Rightarrow \quad Q^x(X_s,a) \approx \hat{r}(X_s,a) + \sum_{t=s+1}^{T-1}\gamma^{t-s}\frac{h_\beta(a|X_s,X_t)}{\pi(a|X_s)}R_t + \gamma^{T-s}\frac{h_\beta(a|X_s,X_T)}{\pi(a|X_s)}V(X_T).$$

Credit Assignment - Conditioning on Future States

$$Q^\pi(x,a) = r(x,a) + \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}\Big[\sum_{k\geq 1}\gamma^k \frac{h_k(a|x,X_k)}{\pi(a|x)} R_k\Big].$$

counterfactual importance sampling

infeasible, time-dependent

$$\Rightarrow \quad A^\pi(x,a) = r(x,a) - r^\pi(x) + \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}\Big[\sum_{k\geq 1}\Big(\frac{h_k(a|x,X_k)}{\pi(a|x)} - 1\Big)\gamma^k R_k\Big]$$

= 0, when irrelevant

$$h_\beta(a|x,y) \stackrel{\text{def}}{=} \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a | X_k = y, k\sim\rho) \quad \text{where} \quad \rho(k) = \beta^{k-1}(1-\beta)$$

Time-independent version

Credit Assignment - Conditioning on Future States

$$h_\beta(a|x,y) \stackrel{\text{def}}{=} \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|X_k = y, k\sim\rho) \quad \text{where} \quad \rho(k) = \beta^{k-1}(1-\beta)$$

Time-independent version

when **β = γ**

$$\Rightarrow \quad A^\pi(x,a) = r(x,a) - r^\pi(x) + \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}\Big[\sum_{k\geq 1}\Big(\frac{h_\beta(a|x,X_k)}{\pi(a|x)} - 1\Big)\gamma^k R_k\Big]$$

PG Algorithm

HCA | State

$$\Rightarrow$$

$$\nabla_\theta V^{\pi_\theta}(x_0) = \mathbb{E}_{\tau\sim\mathcal{T}(x_0,\pi_\theta)}\Big[\sum_{k\geq 0}\gamma^k\sum_a \nabla\pi_\theta(a|X_k)Q^x(X_k,a)\Big]$$

$$Q^x(X_s,a) \approx \hat{r}(X_s,a) + \sum_{t=s+1}^{T-1}\gamma^{t-s}\frac{h_\beta(a|X_s,X_t)}{\pi(a|X_s)}R_t + \gamma^{T-s}\frac{h_\beta(a|X_s,X_T)}{\pi(a|X_s)}V(X_T)$$

Credit Assignment - Conditioning on Future Returns

$$\frac{h(a|x,\pi,f(\tau))}{\pi(a|x)}$$

Predictive Coding

Future Returns

$$h_z(a|x,\pi,z) \overset{def}{=} \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}\left(A_0 = a \middle| Z(\tau) = z\right).$$

Bayes' rule:

$$\frac{\pi(a|x)}{h_z(a|x,\pi,z)} = \frac{\mathbb{P}\left(Z(\tau)=z\right)}{\mathbb{P}\left(Z(\tau)=z|A_t=a\right)} = \frac{\mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}\left(Z(\tau)=z\right)}{\mathbb{P}_{\tau\sim\mathcal{T}(x,a,\pi)}\left(Z(\tau)=z\right)}$$

trajectories start with *x* and *a*

Thm. 2

$$\Rightarrow \qquad V^\pi(x) = \mathbb{E}_{\tau\sim\mathcal{T}(x,a,\pi)}\left[Z(\tau)\frac{\pi(a|x)}{h_z(a|x,Z(\tau))}\right].$$

importance sampling

Credit Assignment - Conditioning on Future Returns

$$V^\pi(x) = \mathbb{E}_{\tau \sim \mathcal{T}(x,a,\pi)}\left[Z(\tau)\frac{\pi(a|x)}{h_z(a|x, Z(\tau))}\right].$$

importance sampling

$$\Rightarrow \quad A^\pi(x,a) = \mathbb{E}_{\tau \sim \mathcal{T}(x,a,\pi)}\left[\left(1 - \frac{\pi(a|x)}{h_z(a|x, Z(\tau))}\right)Z(\tau)\right].$$

"credit" - how much a single action contributed to obtaining a return

credit > 0 if action ***a*** has made achieving ***Z*** more likely

credit < 0 if other actions contributed to achieving ***Z*** more than ***a***

Credit Assignment - Conditioning on Future Returns

$$V^\pi(x) = \mathbb{E}_{\tau \sim \mathcal{T}(x,a,\pi)}\left[Z(\tau)\frac{\pi(a|x)}{h_z(a|x, Z(\tau))}\right].$$

importance sampling

$$\Rightarrow \quad A^\pi(x,a) = \mathbb{E}_{\tau \sim \mathcal{T}(x,a,\pi)}\left[\left(1 - \frac{\pi(a|x)}{h_z(a|x, Z(\tau))}\right)Z(\tau)\right].$$

"credit" - how much a single action contributed to obtaining a return

PG Algorithm

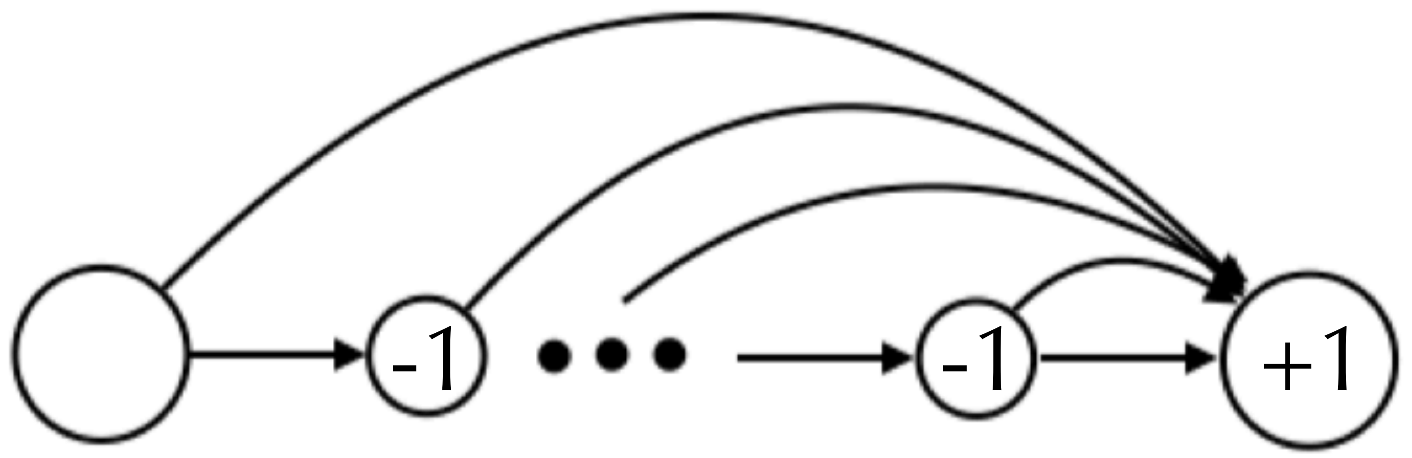
HCA | Return

$$\Rightarrow \quad \nabla_\theta V^{\pi_\theta}(x_0) = \mathbb{E}_{\tau \sim \mathcal{T}(x_0,\pi_\theta)}\left[\sum_{k \geq 0} \gamma^k \nabla \log \pi_\theta(A_k|X_k) A^z(X_k, A_k)\right],$$

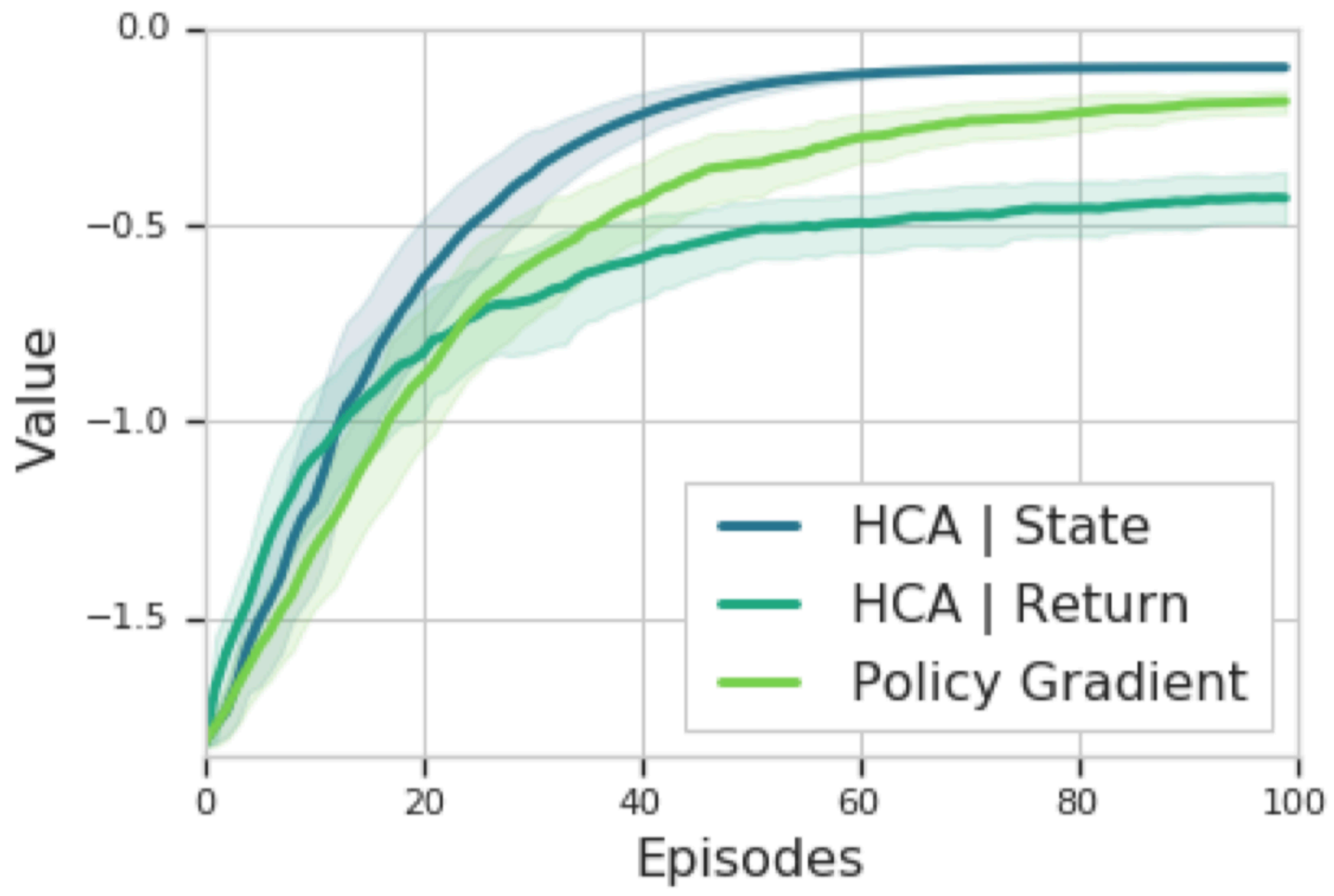
$$A^z(X_s, A_s) = \left(1 - \frac{\pi(A_s|X_s)}{h_z(A_s|X_s, Z_s)}\right) Z_s \quad \text{where} \quad Z_s = \sum_{t \geq s} \gamma^{t-s} R_t.$$

valid "baseline"- even if dependent of actions.

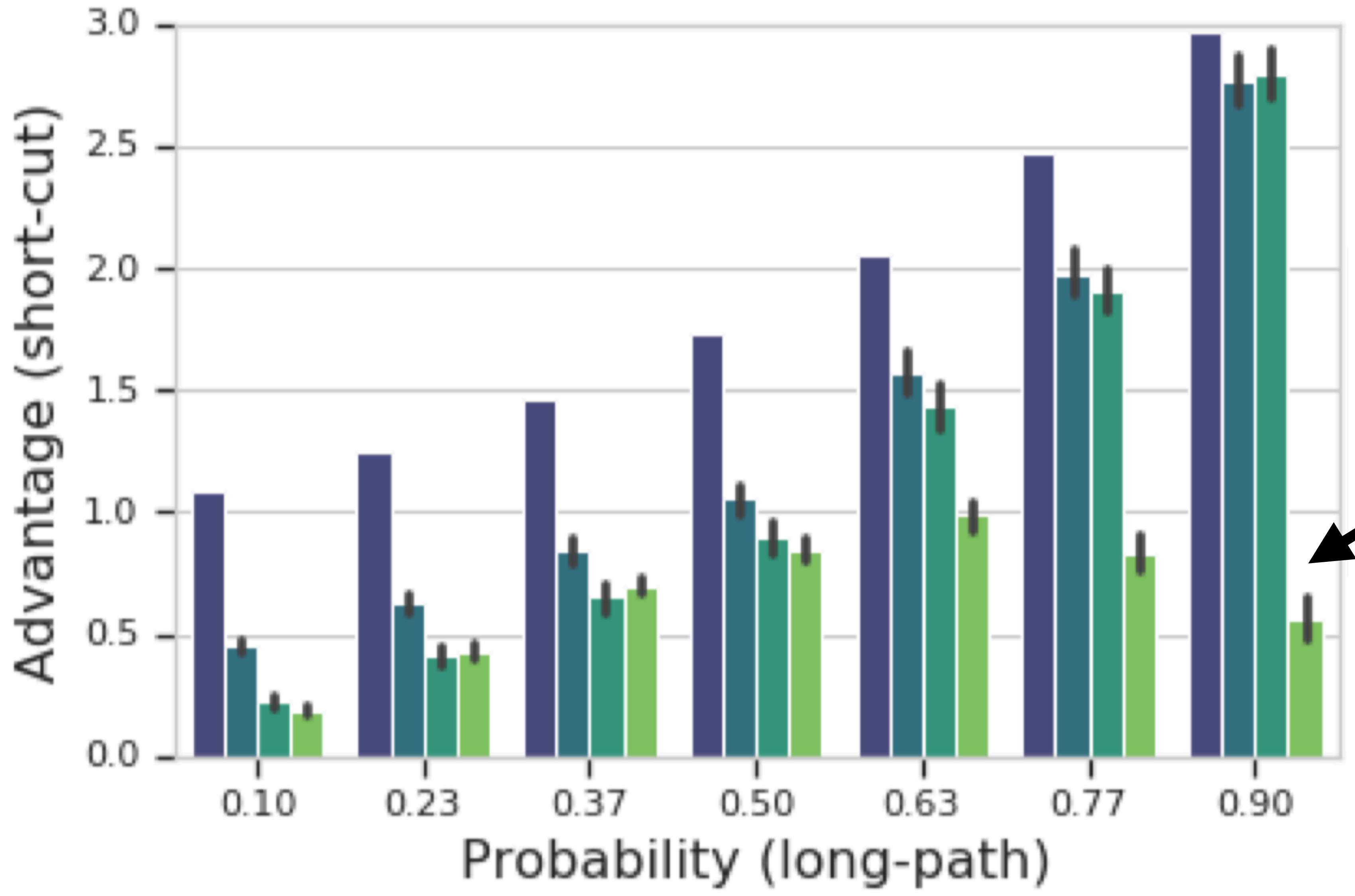
Experiments



Shortcut

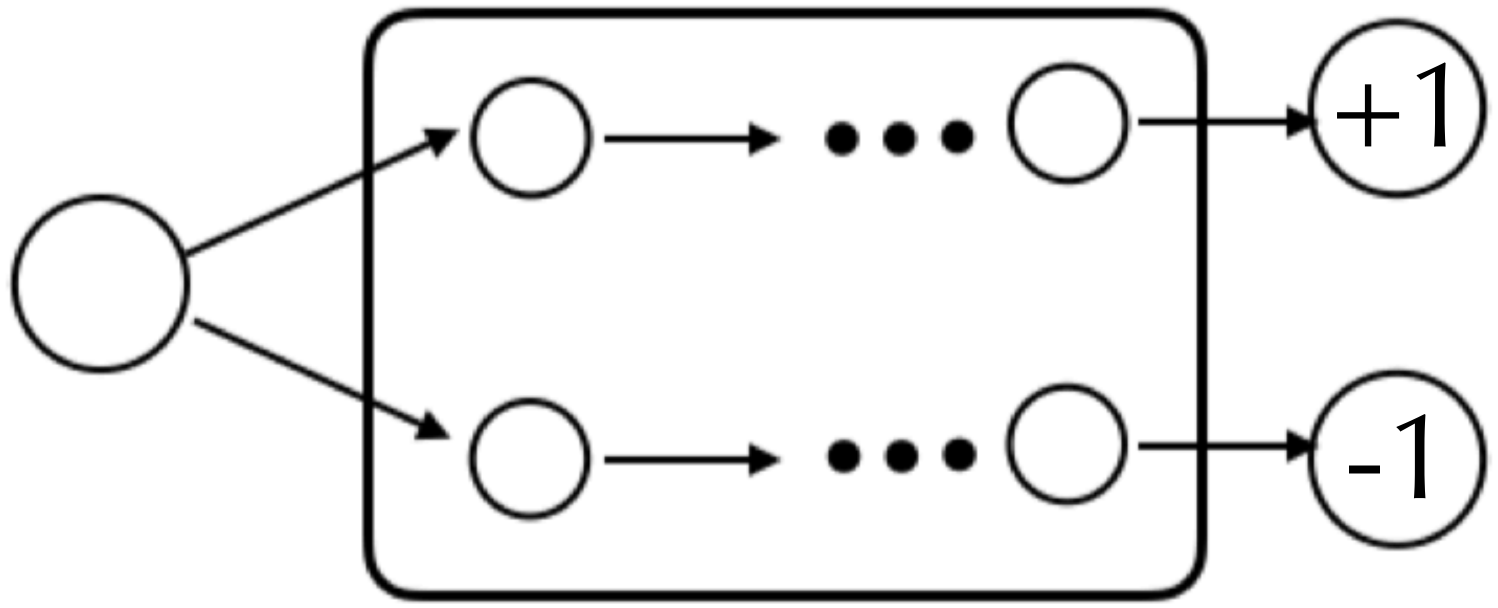


- **counter-factual credit assignment** (issue 4), when the long path is taken more frequently than the shortcut path, counter-factual updates become increasingly effective
- **the use of time as a proxy for relevance (issue 3)** is shown to be only a heuristic, even in a fully-observable MDP.

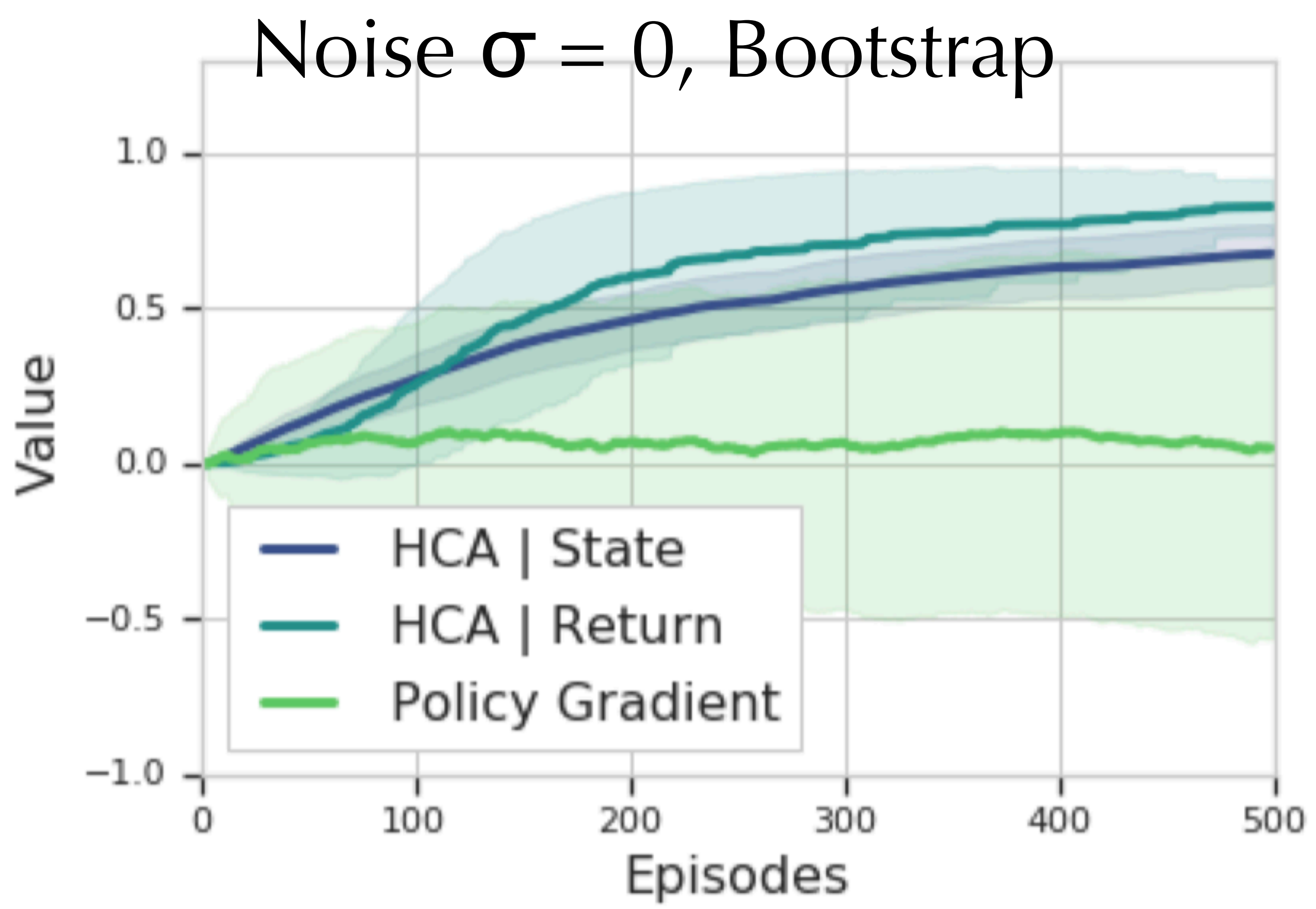


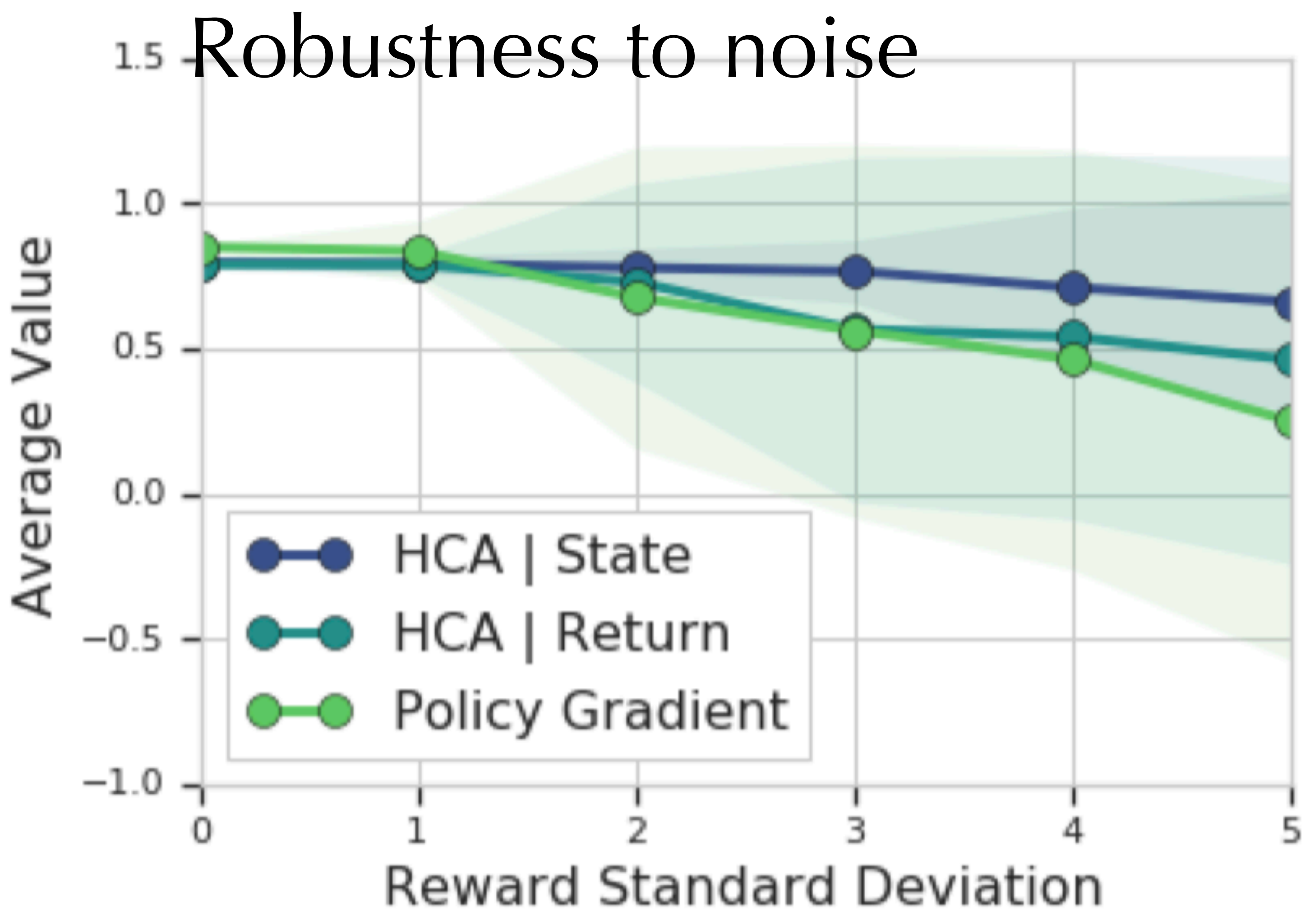
The relevance for the states along the chain is not accurately reflected in the long temporal distance between them and the goal state.

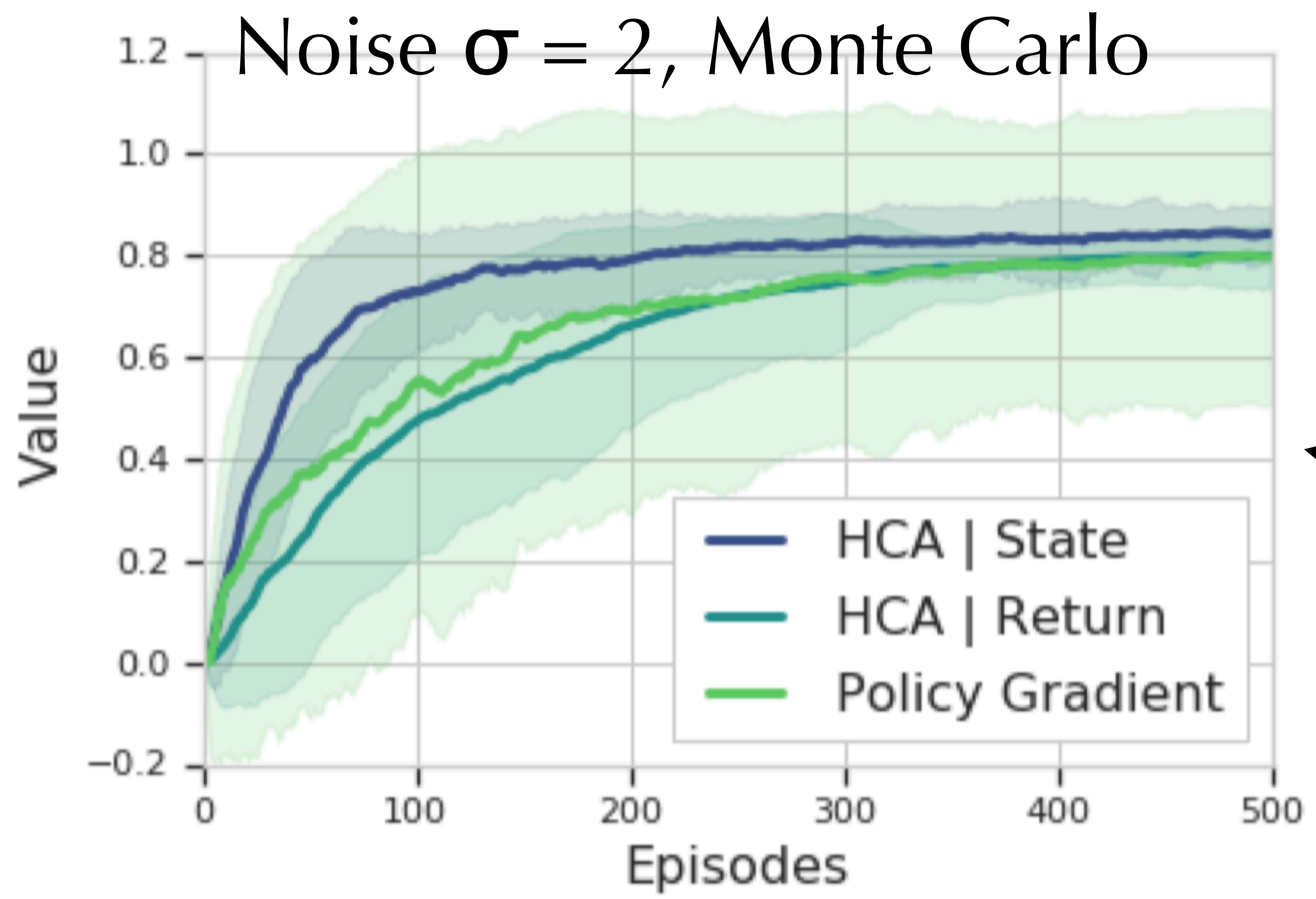
Experiments



Delayed effect.



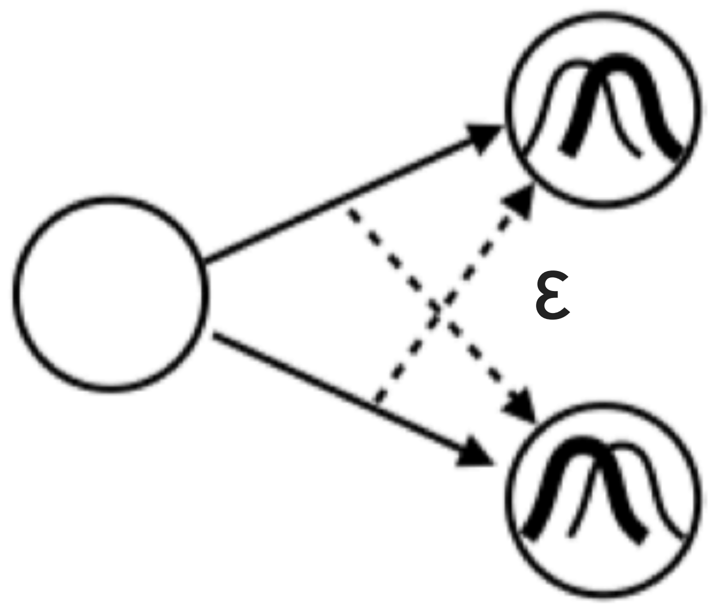




Return-conditional HCA is a harder learning problem: eq. to learning values

- **Bootstrapping naively is inadequate in this case (issue 2),** but HCA is able to carry the appropriate information
- **its performance deteriorates when intermediate reward noise is present (issue 1).** HCA on the other hand is able to reduce the variance due to the irrelevant noise in the rewards.
- **using temporal proximity for credit assignment is a heuristic** (issue 3).

Experiments



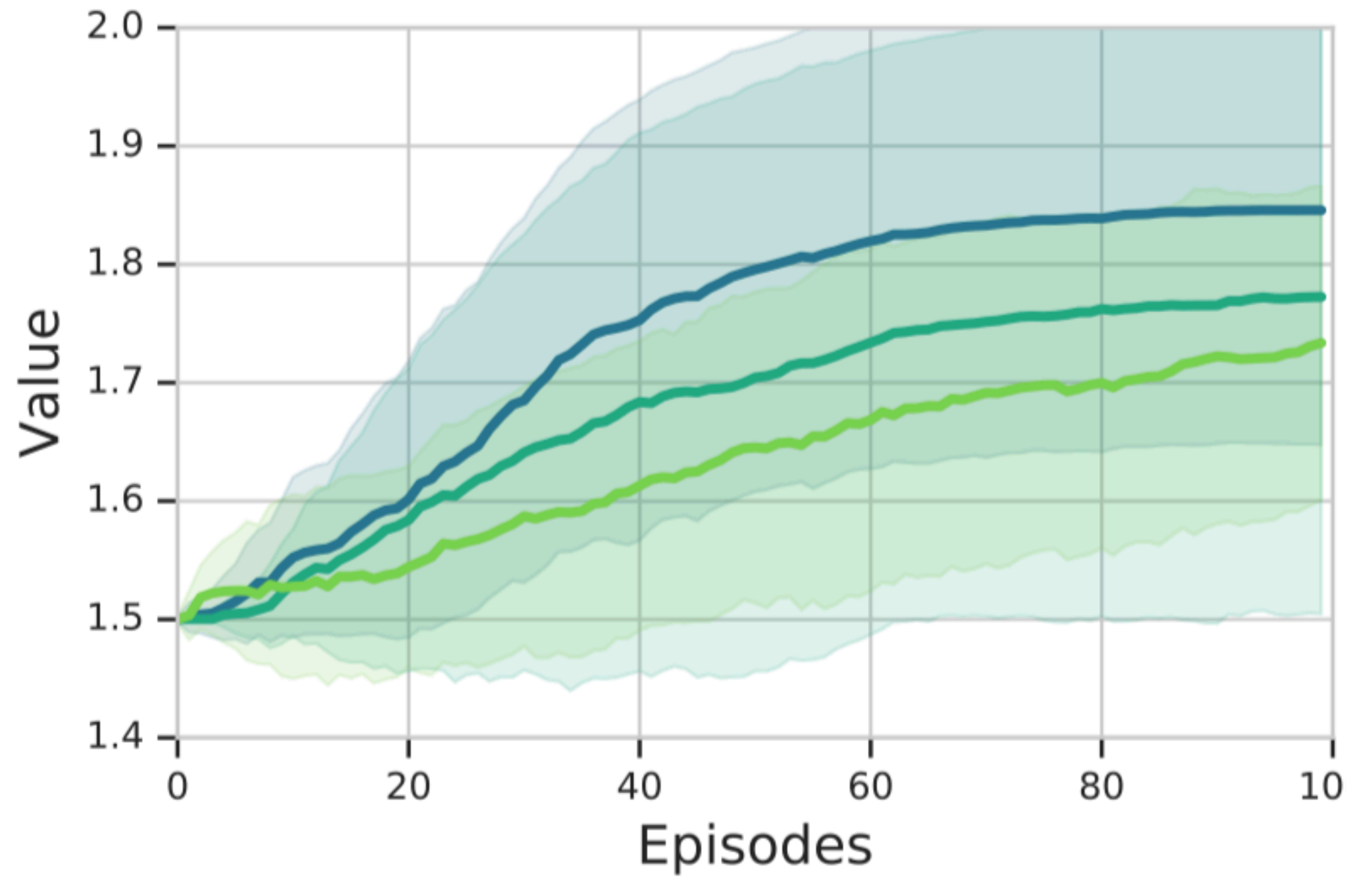
Gaussian $N(1, 1.5)$

Gaussian $N(2, 1.5)$

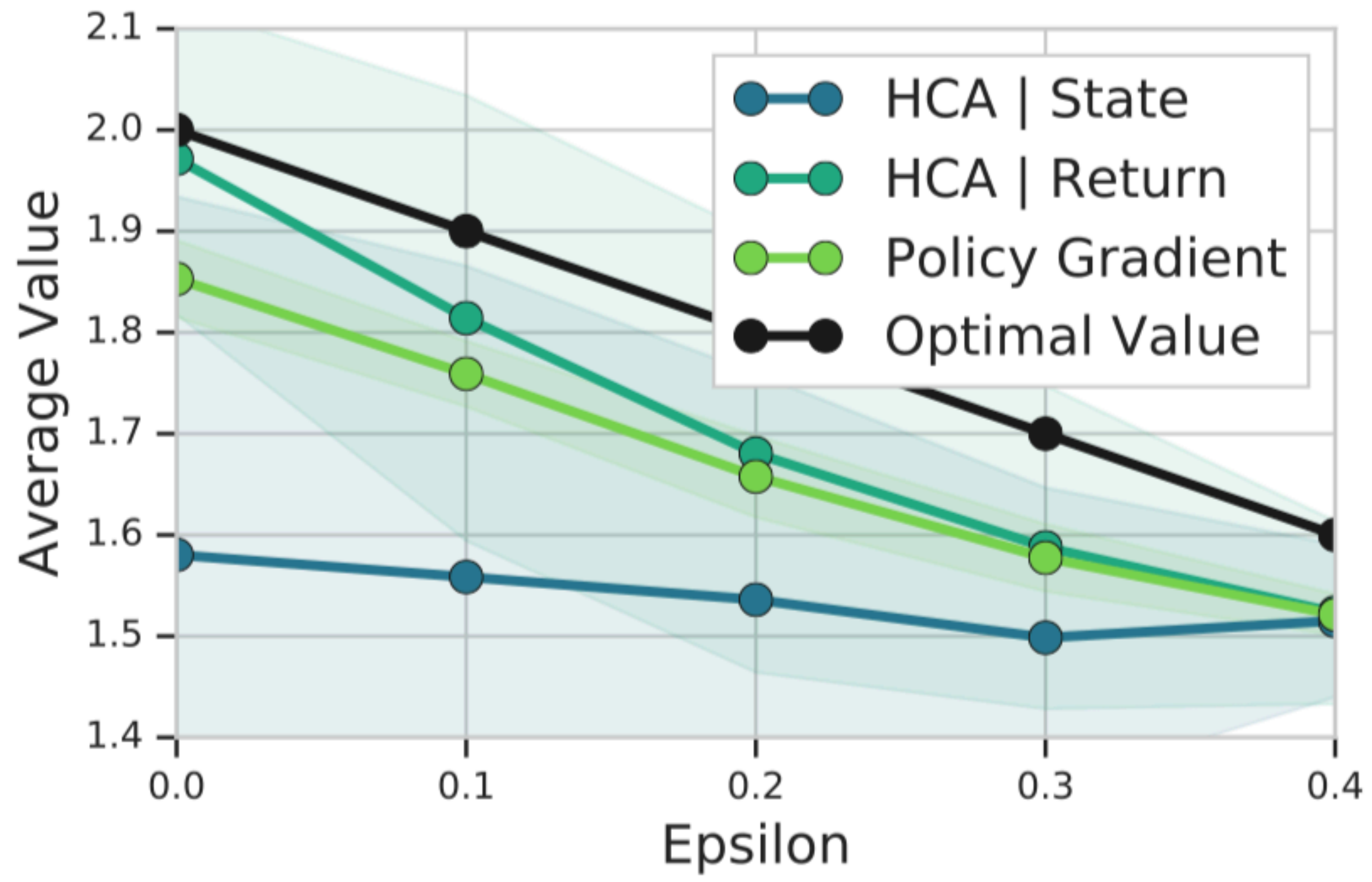
Ambiguous bandit.

- **variance (issue 1)** with some probability **ε** of crossover.
- **a lack of counter-factual updates (issue 4)** difficult to tell whether an action was genuinely better, or just happened to be on the tail end of the distribution.
- **partial observability of the final state (issue 2)**

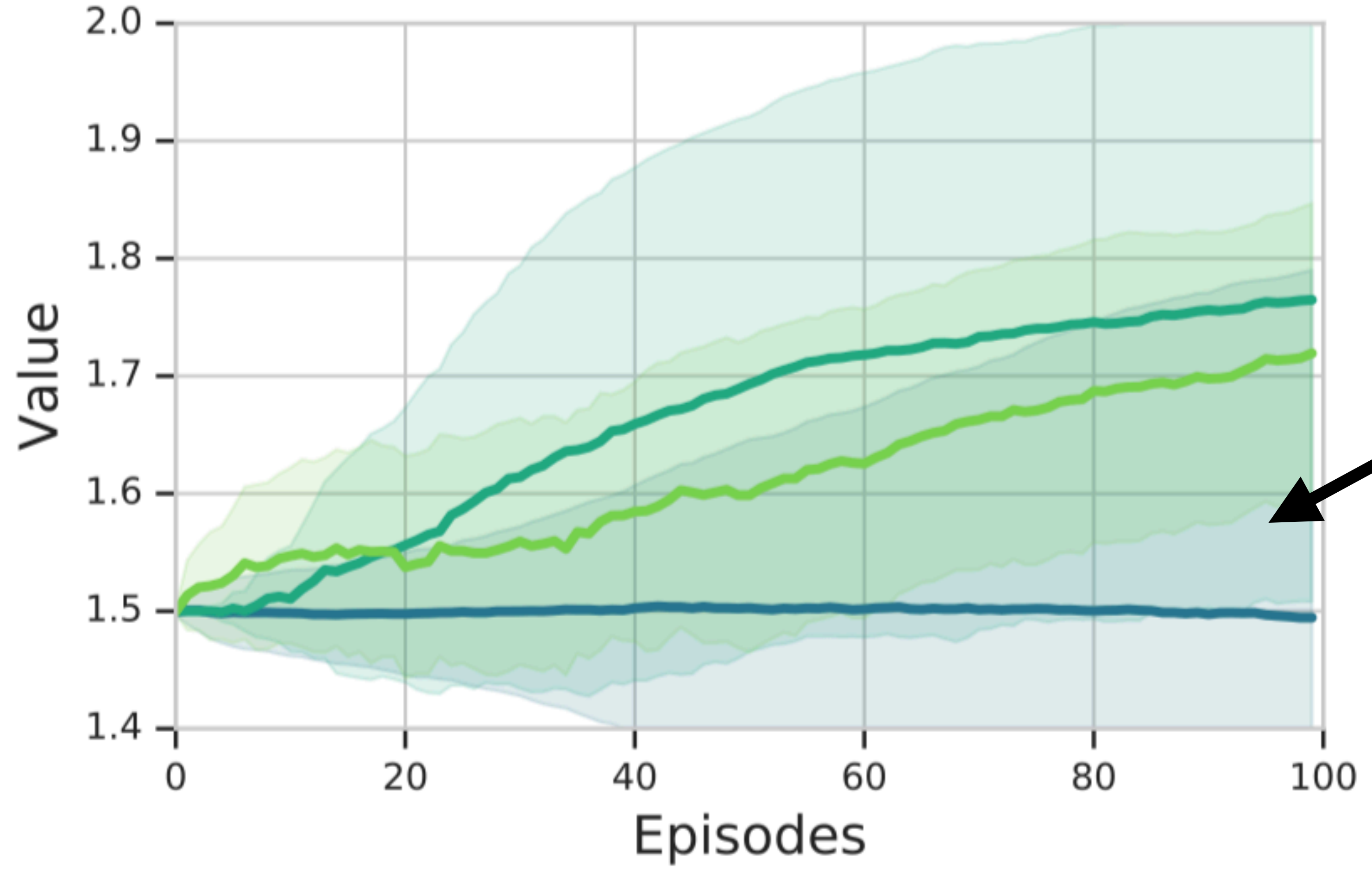
Observable



Hidden state, vary ε, σ = 0.5



Hidden state



Return-conditional policy is still able to improve over policy gradient, but state-conditioning fails.

Hindsight Credit Assignment

$$I(A_t; f(\tau_{t:\infty})|X_t = x) = \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}\left[\log\left(\frac{\mathbb{P}(A = A_t|f(\tau) = f(\tau_{t:\infty}), X_t = x)}{\mathbb{P}(A = A_t|X_t = x)}\right)\right]$$

density ratio depicts relevance of *actions* and *outcomes* given states

$$\frac{h(a|x,\pi,f(\tau))}{\pi(a|x)}$$

Predictive Coding

can be learned by **InfoNCE** and other supervised learning method.

Future States

$$h_k(a|x,\pi,y) \stackrel{def}{=} \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|X_k = y).$$

Future Returns

$$h_z(a|x,\pi,z) \stackrel{def}{=} \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|Z(\tau) = z).$$

Hindsight Credit Assignment

$$I(A_t; f(\tau_{t:\infty})|X_t = x) = \mathbb{E}_{\tau \sim \mathcal{T}(x,\pi)}\left[\log\left(\frac{\mathbb{P}(A = A_t|f(\tau) = f(\tau_{t:\infty}), X_t = x)}{\mathbb{P}(A = A_t|X_t = x)}\right)\right]$$

density ratio depicts relevance of *actions* and *outcomes* given states

Any Theoretical Guarantee or Empirical Evidence of Improvement?

$$\frac{h(a|x,\pi,f(\tau))}{\pi(a|x)}$$

Predictive Coding

can be learned by **InfoNCE** and other supervised learning method.

Future States

$$h_k(a|x,\pi,y) \overset{def}{=} \mathbb{P}_{\tau \sim \mathcal{T}(x,\pi)}(A_0 = a|X_k = y).$$

Future Returns

$$h_z(a|x,\pi,z) \overset{def}{=} \mathbb{P}_{\tau \sim \mathcal{T}(x,\pi)}(A_0 = a|Z(\tau) = z).$$