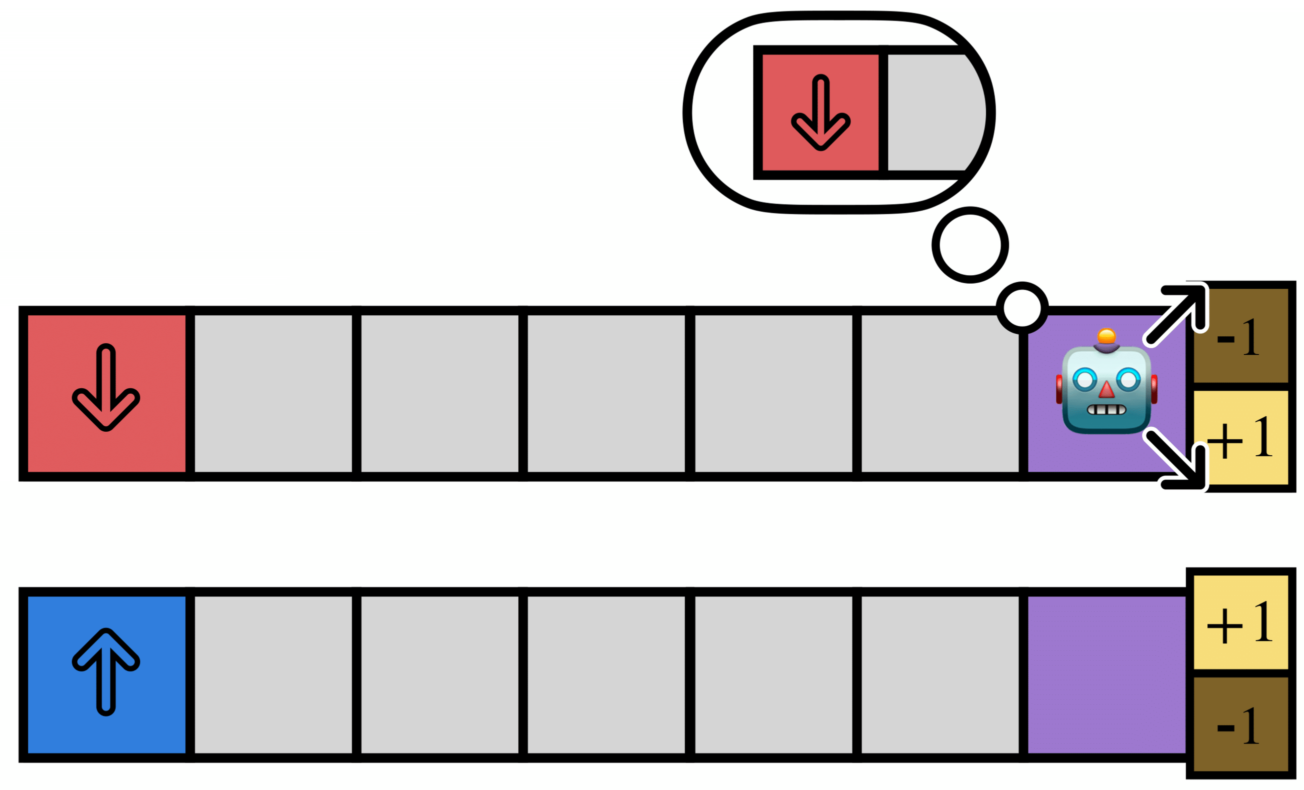
General Value Discrepancies Mitigate Partial Observability in Reinforcement Learning

Peter Koepernik*, Ruo Yu Tao*, Ronald Parr, George Konidaris, Cameron Allen

*Equal Contribution

UNIVERSITY OF OXFORD · BROWN · Duke UNIVERSITY · Berkeley UNIVERSITY OF CALIFORNIA

1 Optimal decision making in partially observable environments requires memory.



2 A memory retains all relevant information if and only if it is *Markov*:

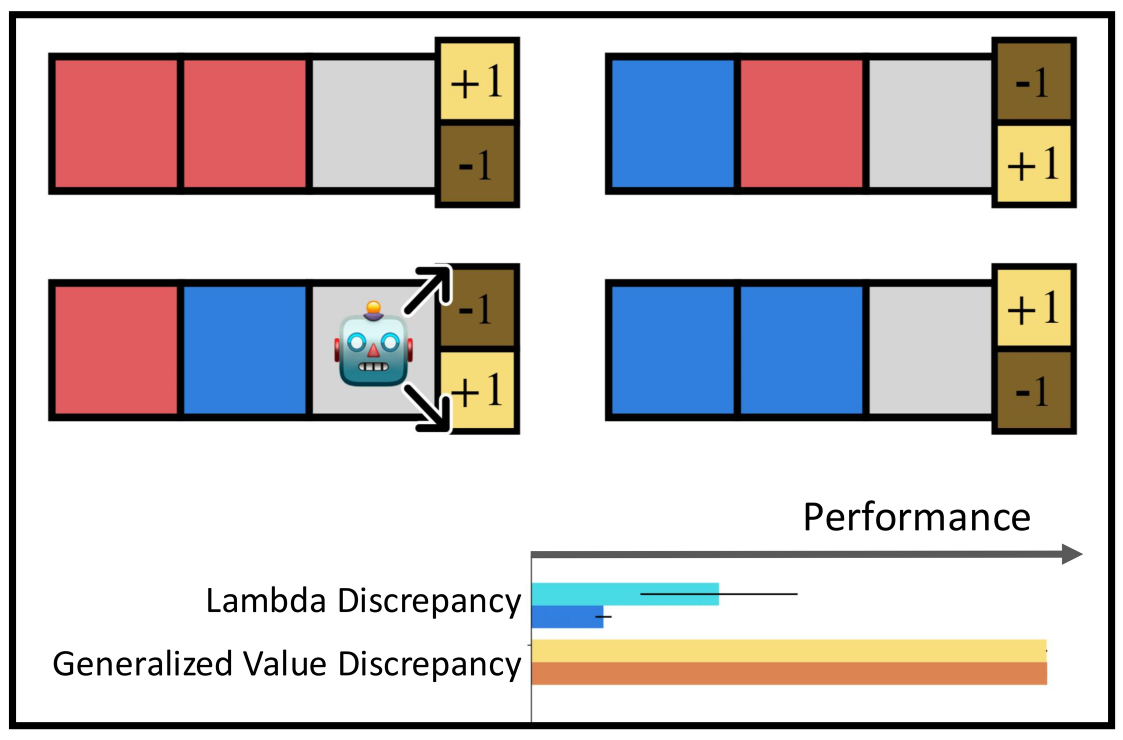
$$\mathbb{P}(\omega_{t+1}, m_{t+1} \mid m_t, \omega_t, \ldots, m_0, \omega_0) = \mathbb{P}(\omega_{t+1}, m_{t+1} \mid m_t, \omega_t)$$

3 If Λ is a measure for "non-Markovianness", we can train the memory m_ϕ with gradient descent.

$$\phi \leftarrow \phi - \alpha \nabla_\phi \Lambda$$

4 **Lambda discrepancy** [1] does this by measuring the difference between MC and TD value estimates: $\Lambda = \left\| V^\pi_{\mathrm{MC}} - V^\pi_{\mathrm{TD}} \right\|$

- fails without rewards
- sometimes fails even with rewards → Example



5 ***Generalized*** **Value Discrepancies can *always* detect partial observability!**

$\underbrace{\text{Generalized Value Discrepancies}}_{\text{Difference between MC/TD estimates of generalized value functions}}$

Difference between MC/TD estimates of *generalized* value functions

$$\sum_t \gamma^t r_t \longrightarrow \sum_t \gamma(\omega_1) \ldots \gamma(\omega_t) f(\omega_t)$$

$\gamma(\omega_1) \ldots \gamma(\omega_t)$: observation-dependent discount

$f(\omega_t)$: Any function of observation ("pseudo-reward")

Setting: POMDP where rewards are part of observation: $r_t = R(\omega_t)$

Definition: A ***generalized value function*** is

$$V_{f,\gamma}(\omega) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma(\omega_0) \ldots \gamma(\omega_{t-1}) f(\omega_t)\right]$$

for functions $f : \Omega \to \mathbb{R}$ and $\gamma : \Omega \to (0,1)$. The associated ***generalized value discrepancy*** (GVD) is

$$\Lambda_{f,\gamma} = \left\| V^{\mathrm{MC}}_{f,\gamma} - V^{\mathrm{TD}}_{f,\gamma} \right\|$$

Theorem: There is partial observability if and only if $\Lambda_{f,\gamma} > 0$ for *some* f, γ

Furthermore, if there is partial observability then $\Lambda_{f,\gamma} > 0$ for *almost all* f, γ

Ablation: The theorem does *not* hold with any of the following restrictions:

- f is reward, γ is constant (Lambda discrepancy)
- f is general, γ is constant
- f is reward, γ is general

[1] Allen et al. "Mitigating Partial Observability in Sequential Decision Processes via the Lambda Discrepancy." *NeurIPS* (2024).