Mitigating Partial Observability in Sequential Decision Processes via the Lambda Discrepancy

BROWN
Berkeley UNIVERSITY OF CALIFORNIA
Duke UNIVERSITY

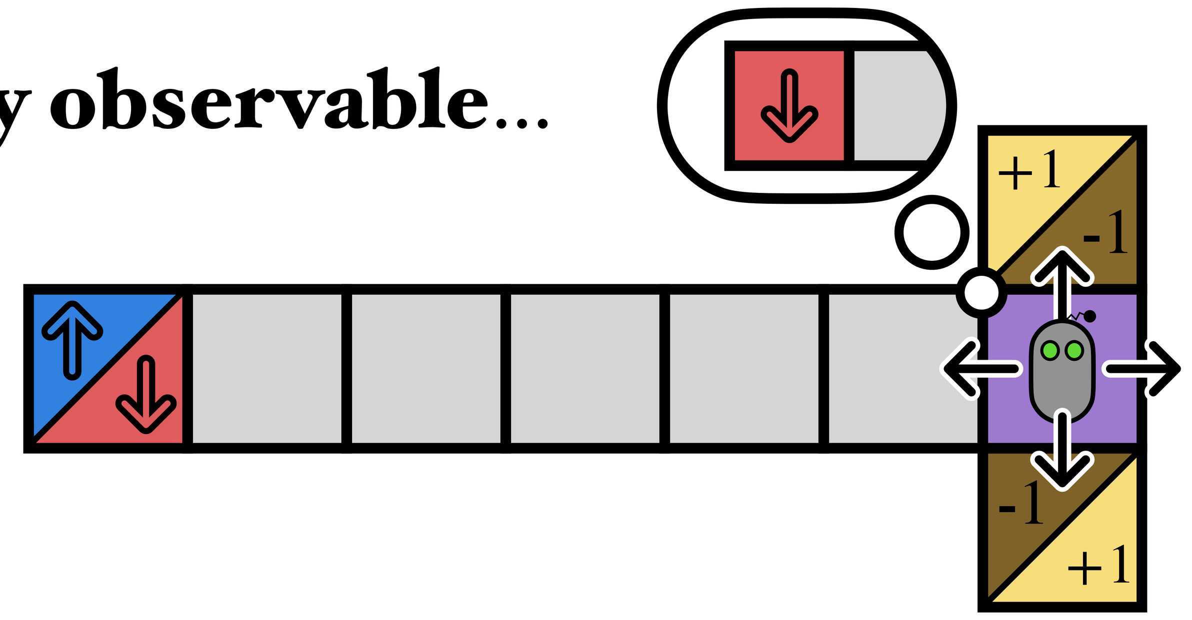
Cameron Allen,* Aaron Kirtland,* Ruo Yu Tao,* Sam Lobel, Daniel Scott, Nicholas Petrocelli, Omer Gottesman, Ronald Parr, Michael L. Littman, George Konidaris

*Equal Contribution

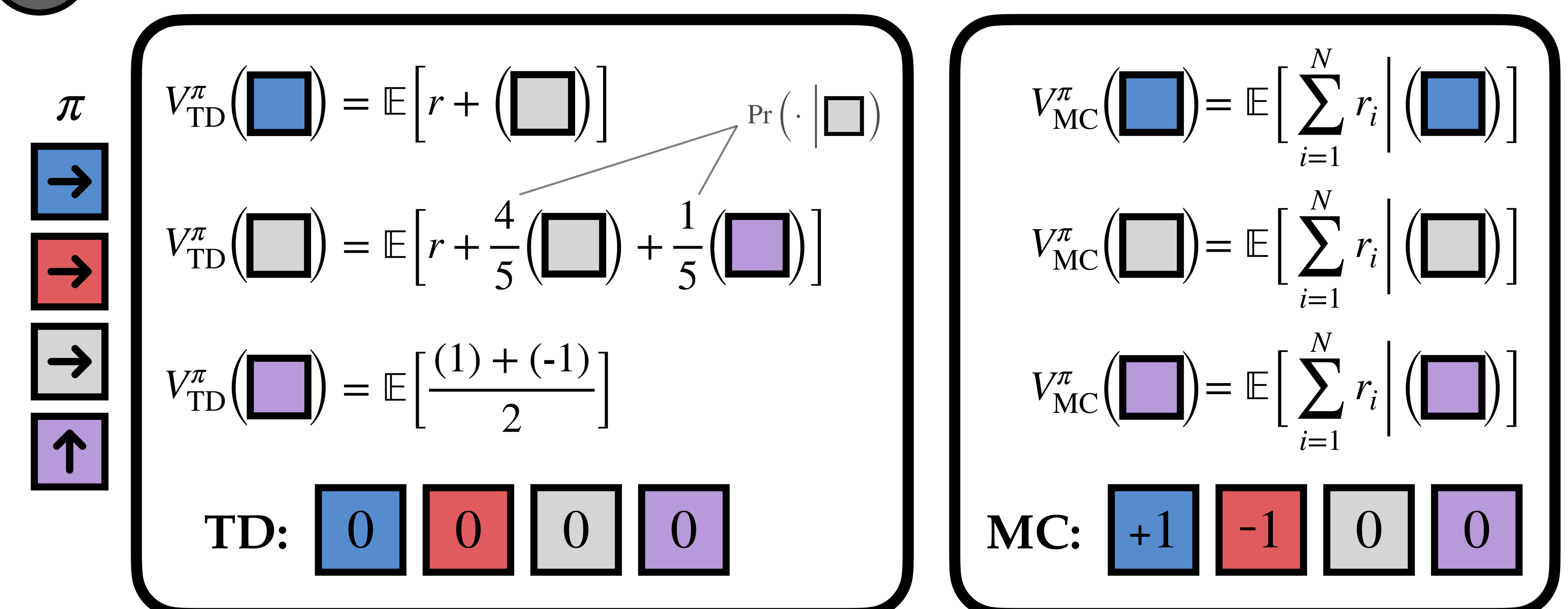
Videos here!

lambda-discrepancy.github.io

1 When the world is **partially observable**...



2 ... **TD** and **Monte Carlo** value estimates can be very different.

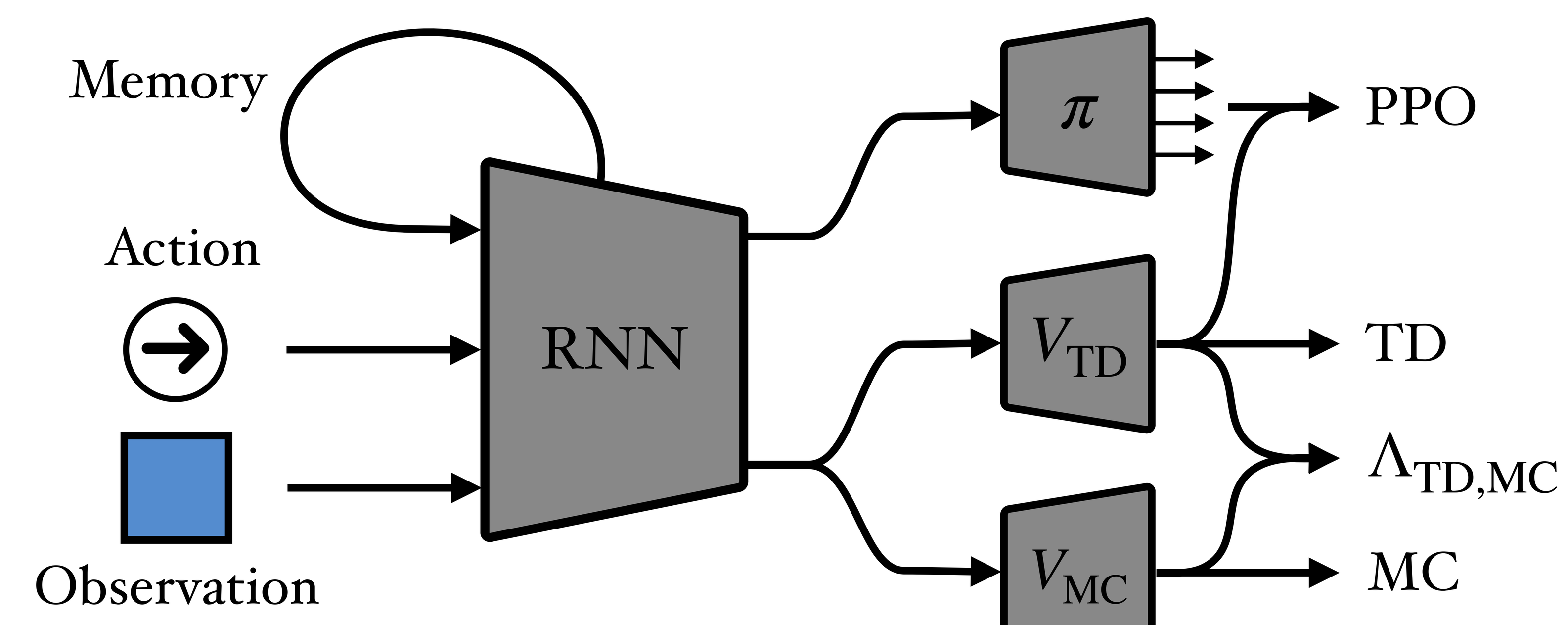


3

$$\Lambda := \left\| V^{\pi}_{TD} - V^{\pi}_{MC} \right\|$$

Reducing value discrepancies helps with learning memory.

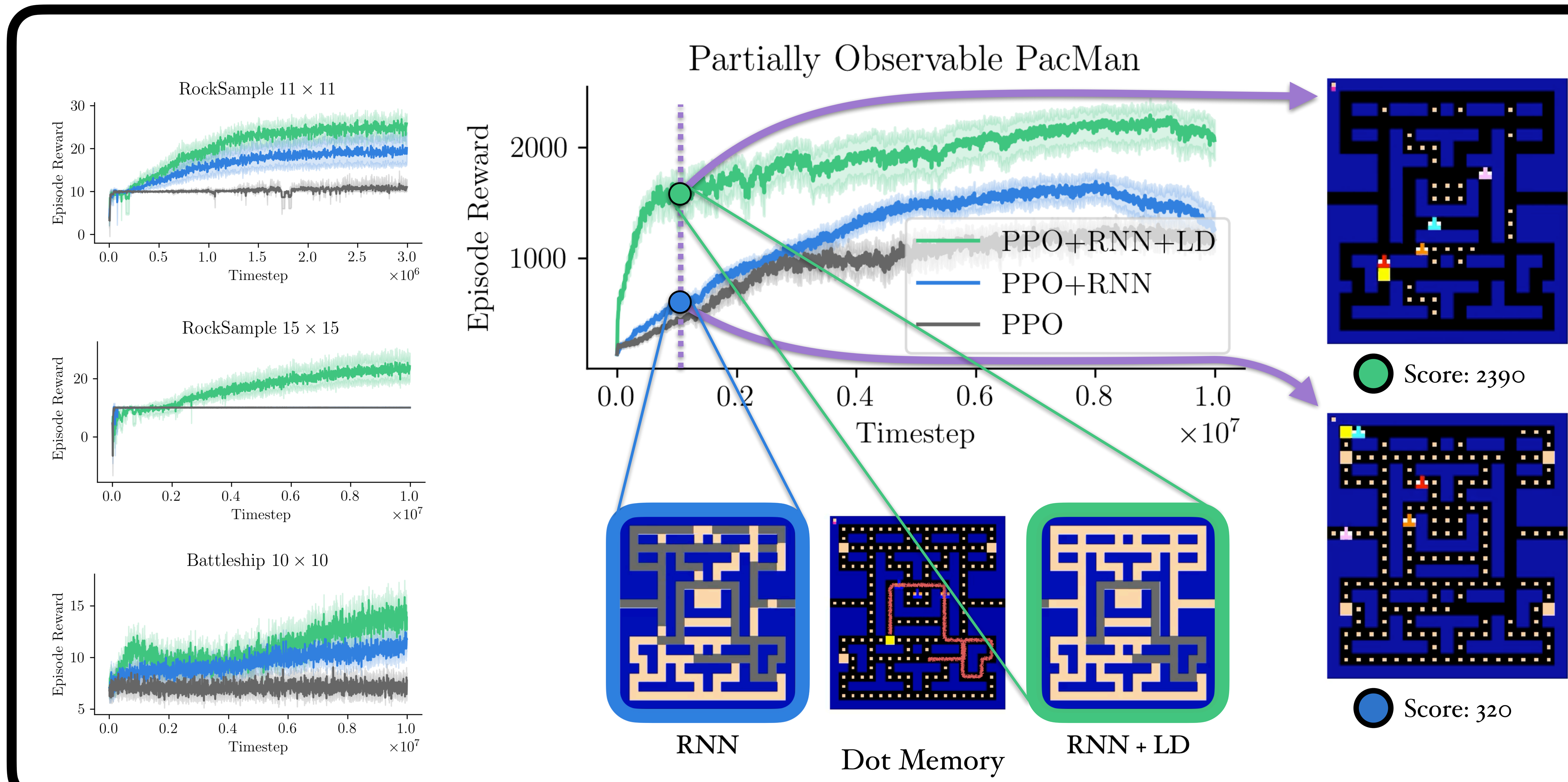
4 Training **TD and MC value functions**, and **minimizing their difference**...



$$\mathcal{L}_{Actor} = \mathcal{L}_{PPO}$$

$$\mathcal{L}_{Critic} = \beta \cdot \Lambda + (1-\beta) \cdot (\mathcal{L}_{TD} + \mathcal{L}_{MC})$$

5 ... leads to **memories** that support **better policies**.



Math Details

TD(λ) blends between TD & MC

$$V^{\lambda}_{\pi_S}(s) = \mathbb{E}_{\pi_S}\left[(1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} g_{t:t+n} \,\middle|\, s_t = s\right]$$

where $g_{t:t+n} := r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^n V_{\pi_S}(s_{t+n})$.

1-step TD (λ=0): $V^{\lambda=0}_{\pi_S}(s) = \mathbb{E}_{\pi_S}[r_t + \gamma V^{\lambda=0}_{\pi_S}(s_{t+1}) \mid s_t = s]$

Monte Carlo (λ=1): $V^{\lambda=1}_{\pi_S}(s) = \mathbb{E}_{\pi_S}[g_t \mid s_t = s]$

TD(λ) varies for POMDPs

$$V^{\lambda=0}_{\Omega}(\omega) = \sum_{a\in A} \pi(a\mid\omega)\Big(R_{\Omega}(\omega,a) + \gamma \sum_{\omega'\in\Omega} T_{\Omega}(\omega'\mid a,\omega) V^{\lambda=0}_{\Omega}(\omega')\Big)$$

$$V^{\lambda=1}_{\Omega}(\omega) = \mathbb{E}_{\pi}\left[g_t \mid \omega_t = \omega\right] = \sum_{s\in S} \Pr(s\mid\omega) V^{\lambda=1}_{S}(s)$$

$$Q^{\lambda}_{\pi} = W\left(I - \gamma T K^{\lambda}_{\pi}\right)^{-1} : R^{SA}$$

where $K^{\lambda}_{\pi} = \left(\lambda \Pi^S + (1-\lambda)\,\Phi W^{\Pi}\right)$

λ-discrepancy can detect POMDPs

$$\Lambda^{\lambda_1,\lambda_2}_{\pi} := \left\| Q^{\lambda_1}_{\pi} - Q^{\lambda_2}_{\pi} \right\| = \left\| W\left(\left(I-\gamma T K^{\lambda_1}_{\pi}\right)^{-1} - \left(I-\gamma T K^{\lambda_2}_{\pi}\right)^{-1}\right) : R^{SA} \right\|$$

Theorems: Given a POMDP $\mathscr{P}$ and distinct $\lambda_1 \neq \lambda_2$ ...

1. If some $\pi : \Omega \to \Delta A$ has $\Lambda > 0$, almost all policies have $\Lambda > 0$.
2. $\left(K^{\lambda_1}_{\pi} = K^{\lambda_2}_{\pi}\right)$ if and only if $\mathscr{P}$ is a block MDP.
3. If $\left(TK^{\lambda_1}_{\pi} = TK^{\lambda_2}_{\pi}\right)$ then $\mathscr{P}$ is equivalent to an MDP.

Remark: For MDPs, $\Lambda = 0$.