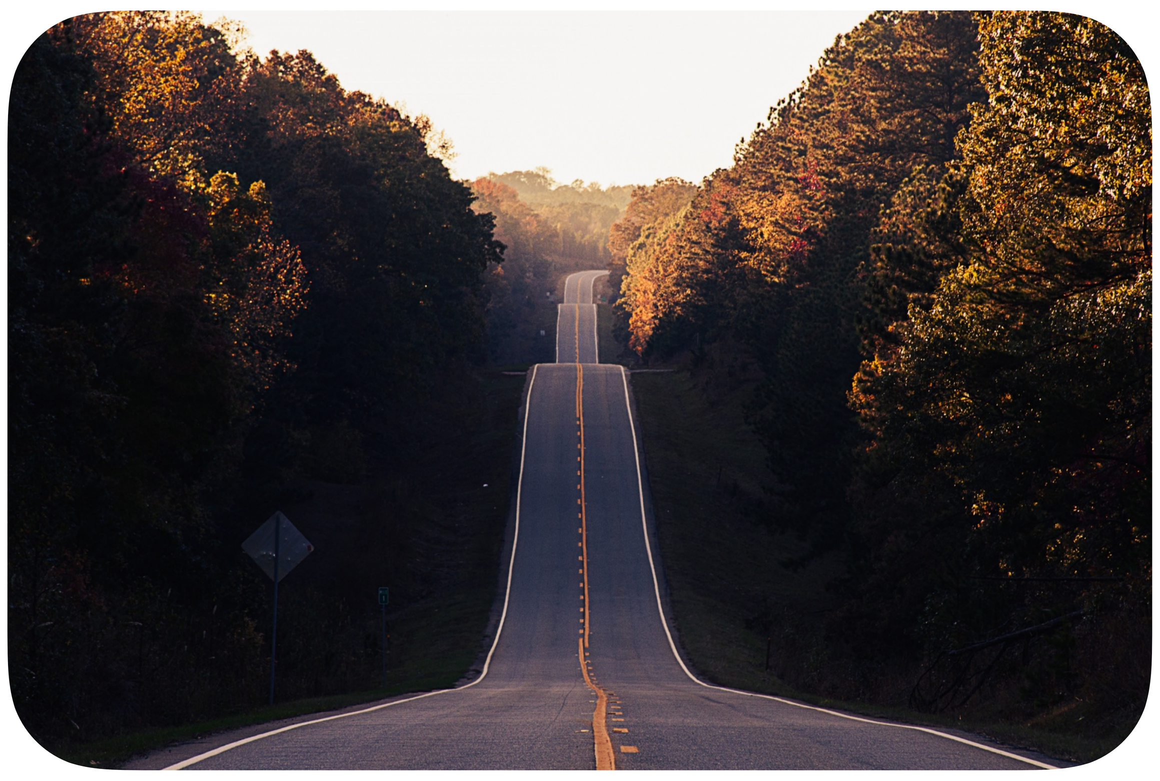
Discounted Reinforcement Learning Is Not an Optimization Problem

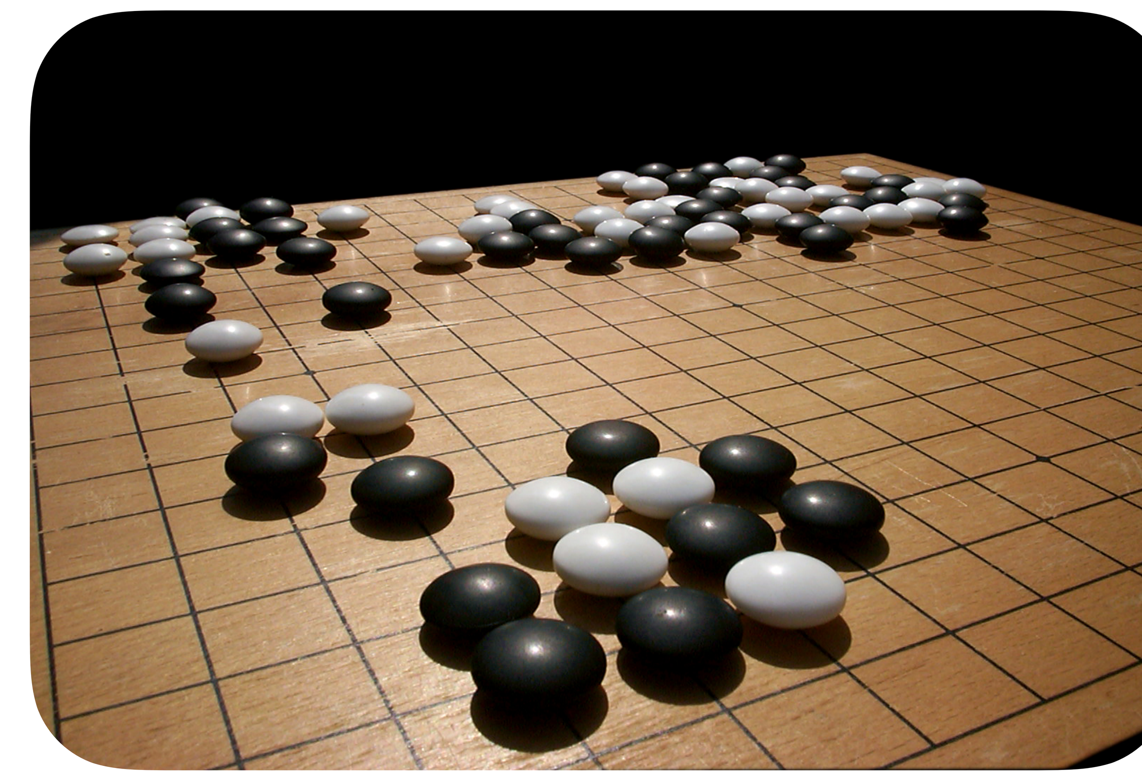
Abhishek Naik   Roshan Shariff   Niko Yasui   Hengshuai Yao   Richard S. Sutton
{anaik1, rshariff, yasui, rsutton}@ualberta.ca   hengshuai.yao@huawei.com
UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Continuing Control



No episodes!
Life goes on forever

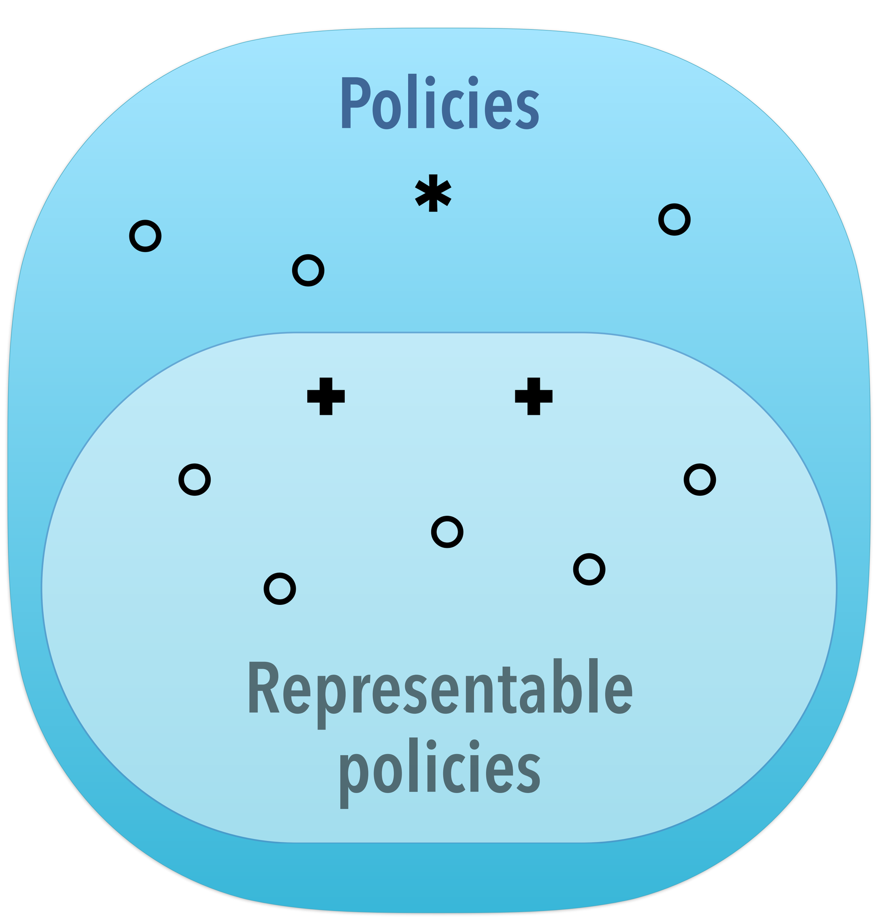
Function Approximation



10^{172} states!
Need to approximate

Continuing control + function approximation → no discounting in the objective

Not an Optimization Problem



Comparing policies:

$$v_\pi^\gamma(s) \geq v_{\pi'}^\gamma(s) \qquad \forall s$$

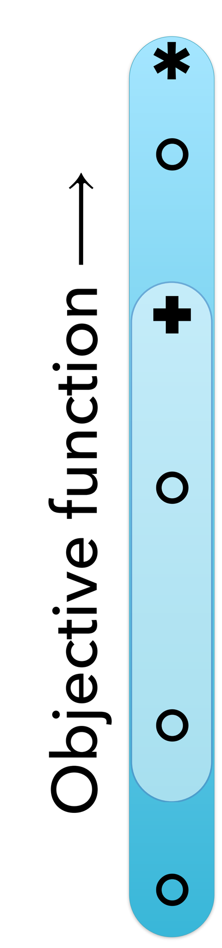
No representable policy is better for every state!

- ○ Policy
- ✱ Optimal policy (for every state)
- ✚ Many "optimal" representable policies (each better in some states, worse in others)

Need an Objective Function

How to evaluate policy performance:

- **Value of start state?**
  - ✔ Episodic tasks
  - ✘ Continuing tasks (start states do not matter)
- **Weighted average over states?**
  - ✔ Average reward
  - ✔ Interest function (but changes problem formulation)

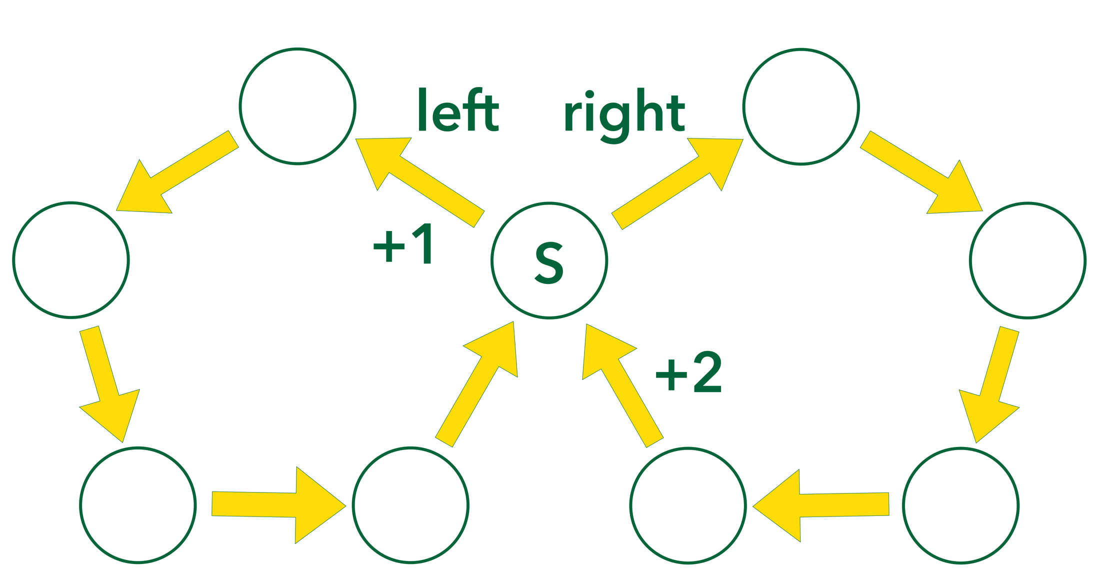


Naively using discounted algorithms should not be the first choice for continuing tasks

✘ Maximizing Discounted Return ✘

Greedily maximizing discounted value does not maximize average reward

(Sarsa, Q-Learning)



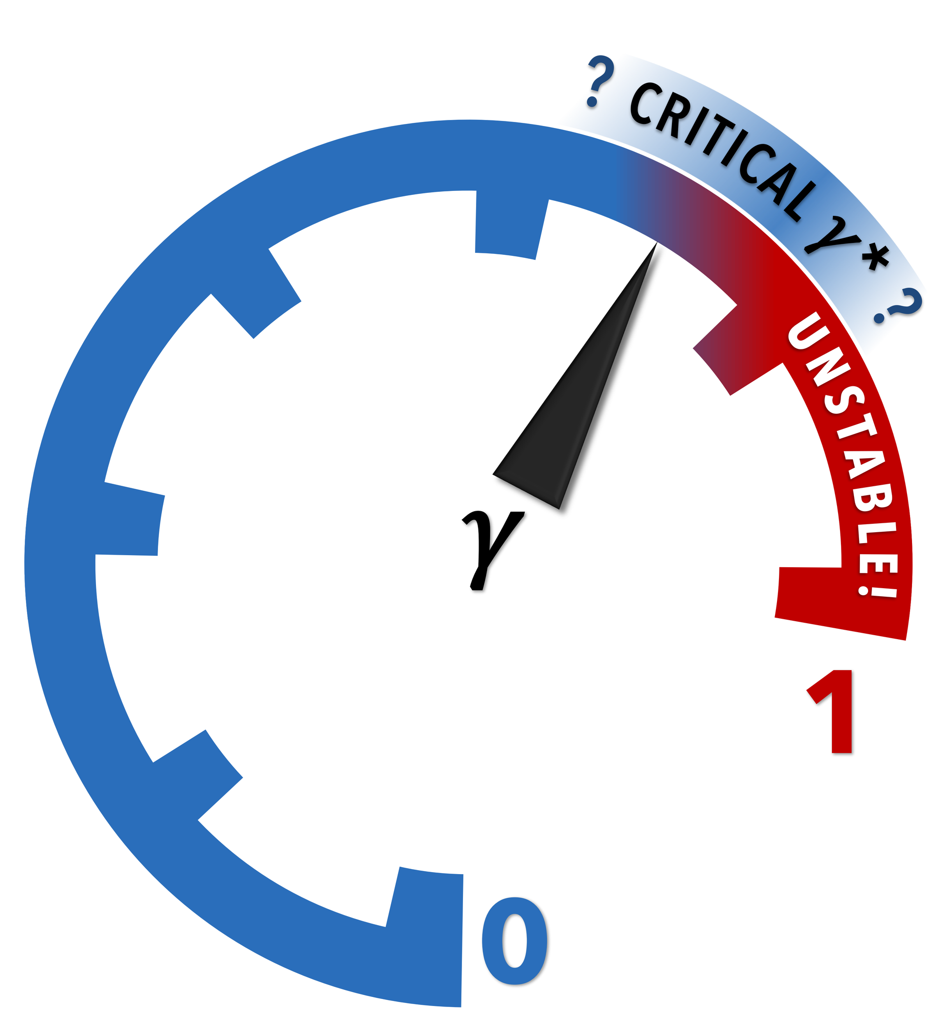
Optimal policy depends on γ

Small γ: go left

Critical γ^* ≈ 0.84

Large γ: go right

✘ Increasing $\gamma \to 1$ ✘



Critical γ is unknown and problem-specific

Algorithms become unstable as $\gamma \to 1$

See paper for details: arxiv.org/abs/1910.02140