Actor-Critic based Improper Reinforcement Learning

Mohammadi Zaki (IISc), Avinash Mohan (Boston Univ.),
Aditya Gopalan (IISc), Shie Mannor (Technion/NVIDIA)

39th International Conference on Machine Learning

Motivation 1 – ‘Simple to Complex’

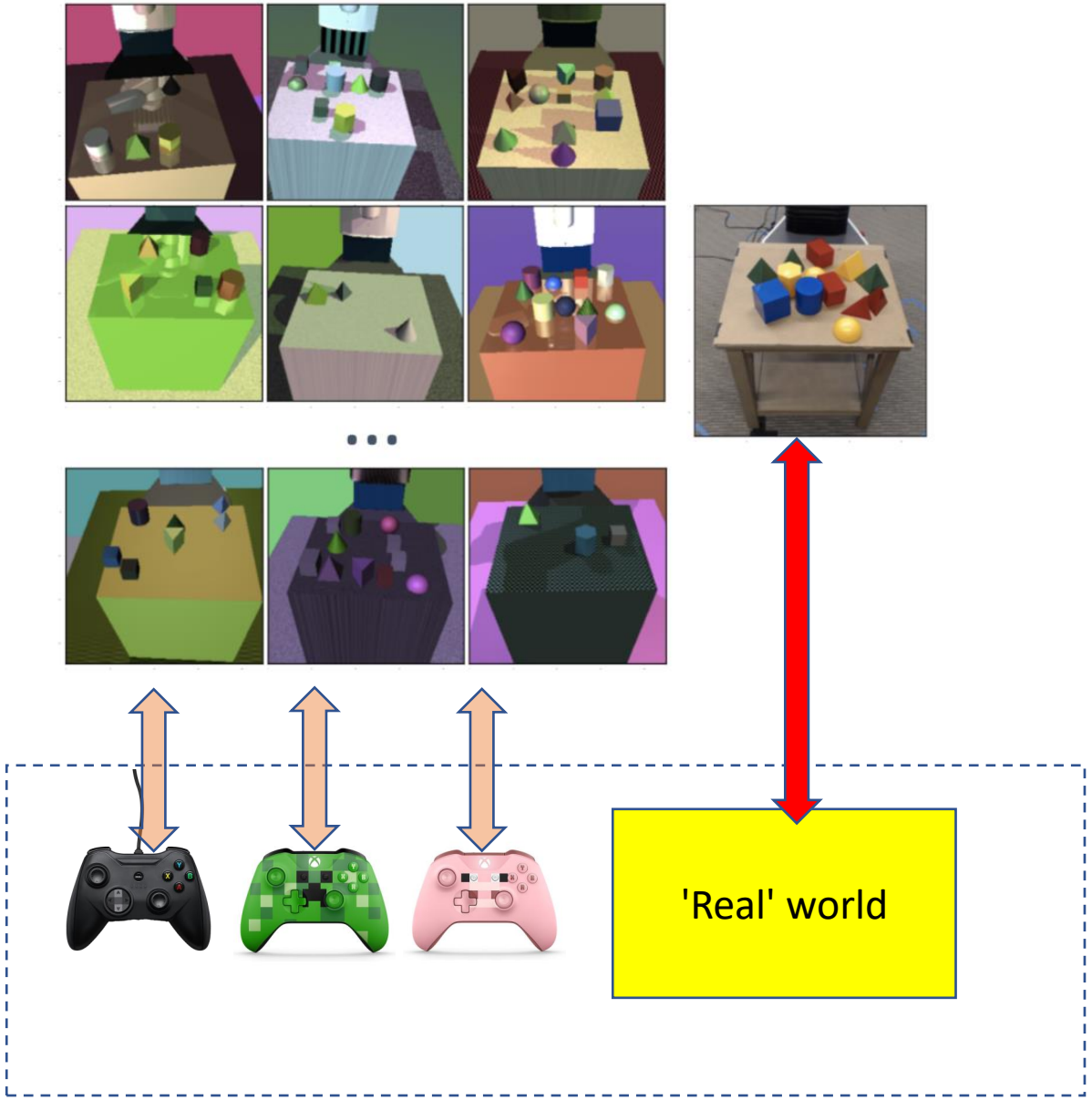
- Complex systems
  - Large (or ∞) # of states, actions
  - Complicated dynamics

- Hierarchical approach to optimal control:
  1. Take 'base controllers' designed to be okay (not great), but
     1. Interpretable
     2. Hand-designed
     3. 'Principled'
     4. Safe
     5. Reliable
  2. Learn to combine them to get something much better ?

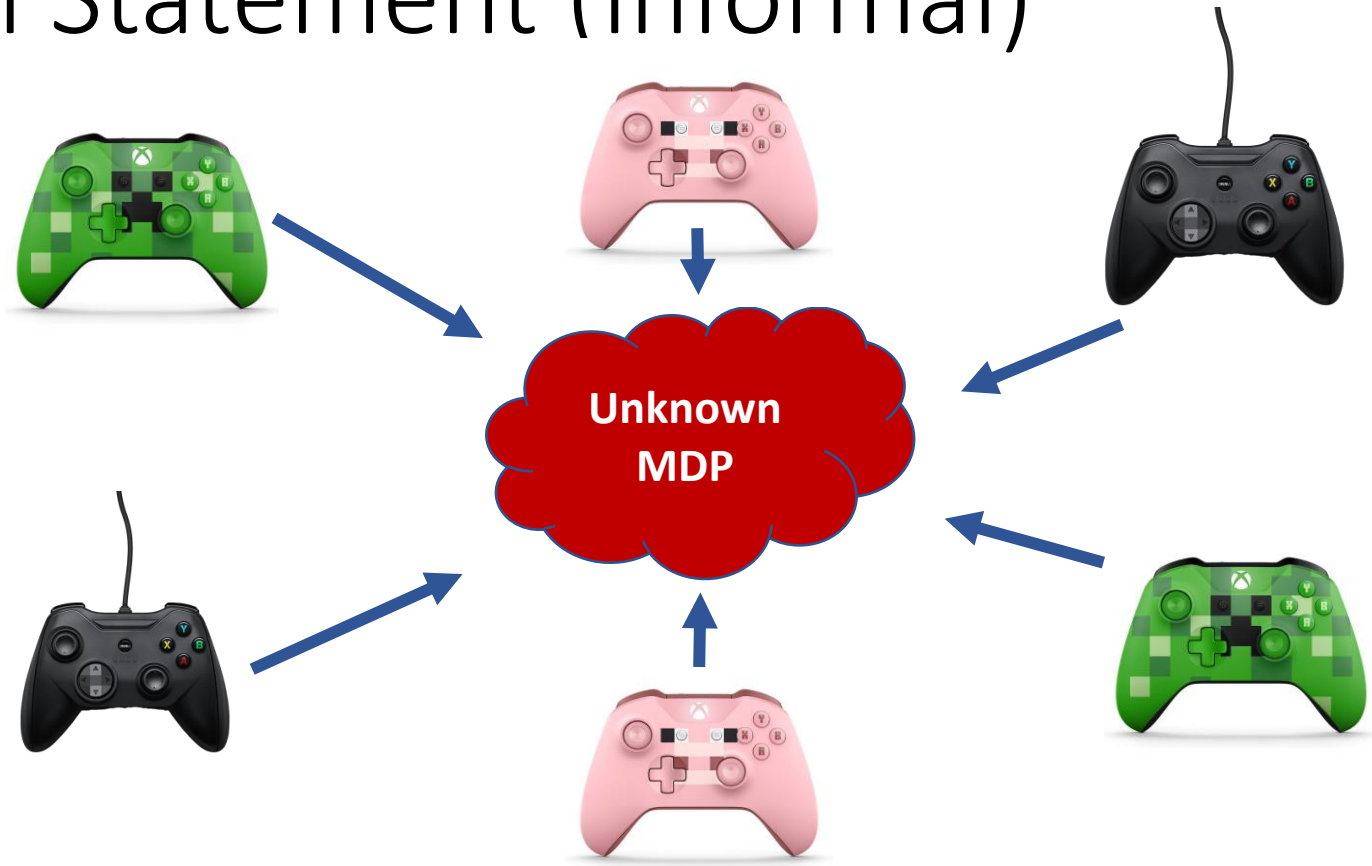
`Simple to Complex’ controller

Motivation 2 – Sim2Real

- Computer simulation enables ‘optimal’ controllers
- But real world may be quite different (Sim2Real gap)
- Approach:
  1. Learn good controllers for a range of simulated environments
  2. (Learn to) Combine them with limited access to the real world



Problem Statement (Informal)



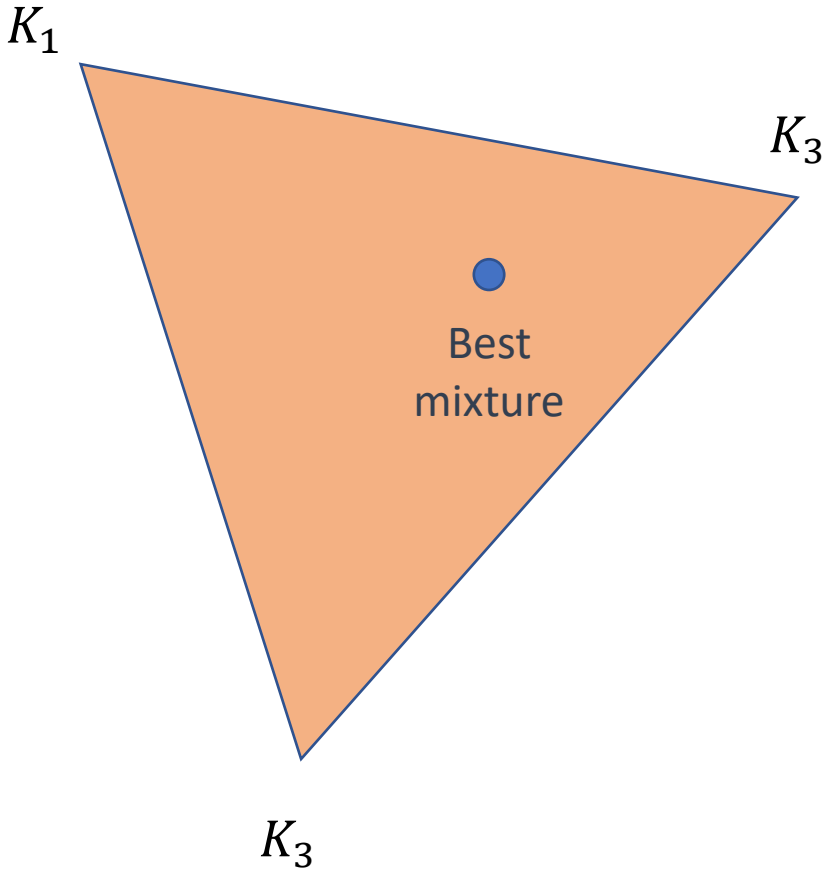
Q: Given a set of custom-made controllers, can we obtain a good (stable, interpretable, etc.) controller for a given (unknown) target environment with relatively few trials?

Building Up

- Given a finite set $\mathcal{C} := \{K_1, \ldots, K_M\}$ of 'base controllers'
  - Each controller $K_i : \mathcal{S} \to \Delta_{\mathcal{A}}$

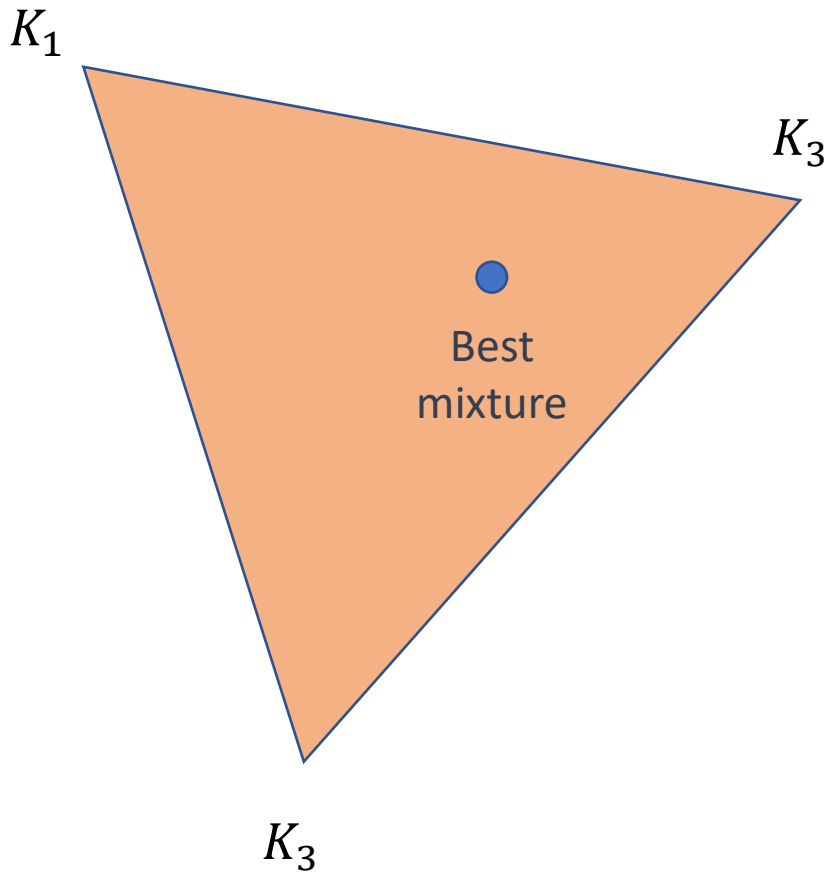
- Improper (mixture) policy class
  - Each point in this class is a probability distribution over the 'base controllers'

- We propose gradient ascent over this simplex

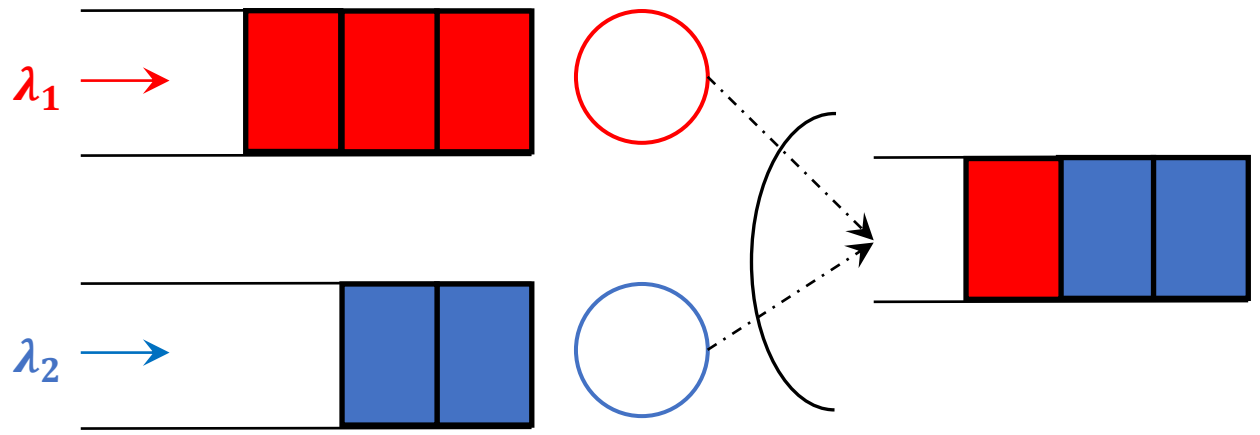


Algorithms & Theoretical Results

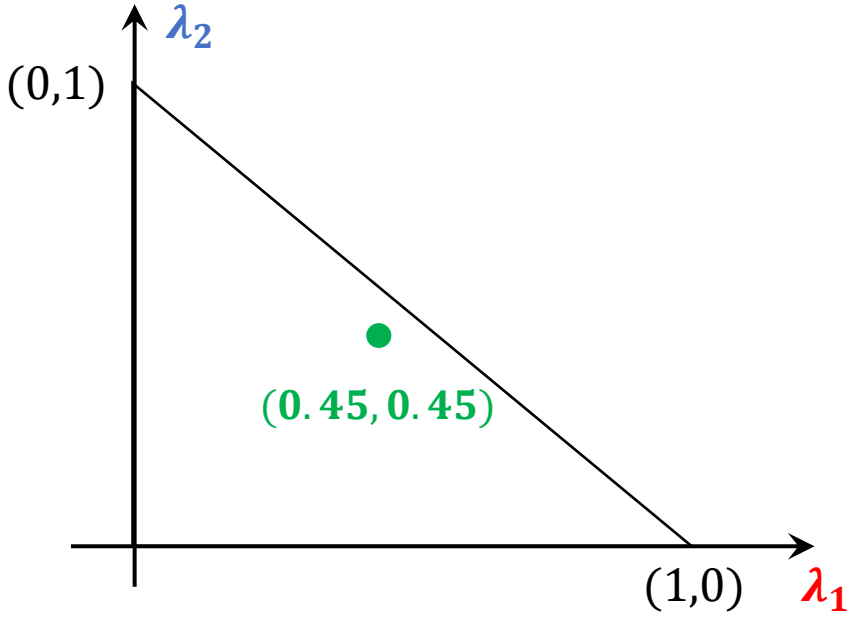
1. Policy-Gradient based
    1. Softmax PG
        - Convergence (with rates) to global optima
2. Actor-Critic (AC) based
    1. Standard AC (With 1st order gradients)
        - Convergence to stationary point
    2. Natural AC (With 2nd order gradients)
        - Convergence to global optima



Simulation on Queue Scheduling- Setting

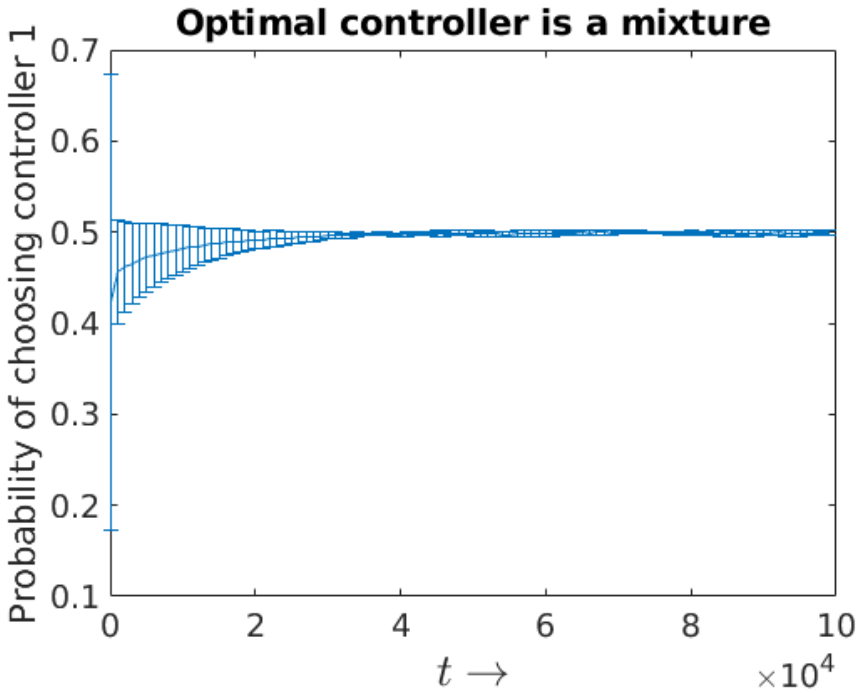


$\leq$ **1 queue**

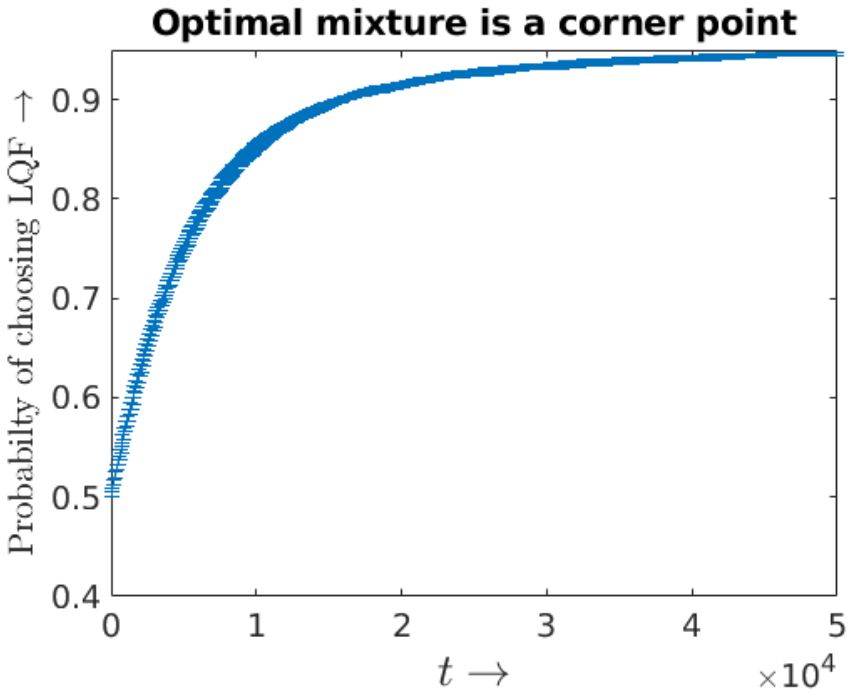


Simulation on Queue Scheduling -- NACIL

Base controllers: {Serve only queue i}, i={1,2}



Base controllers: {{Serve only queue i}, i={1,2}, LQF}

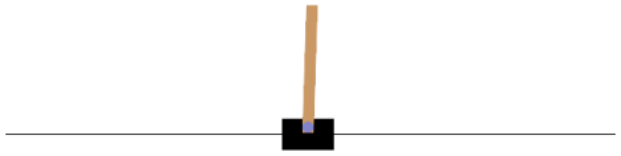


$$(\lambda_1, \lambda_2) \equiv (0.45, 0.45)$$

Simulation on Cartpole-v1 – Softmax PG

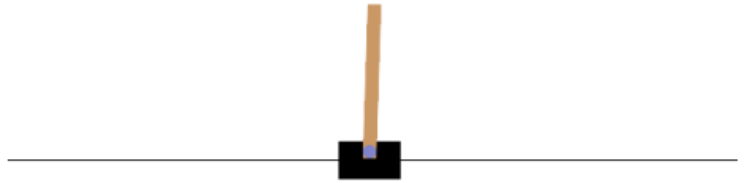
$K_{opt} + \Delta$

$K_{opt} - \Delta$



Base controllers

Softmax PG

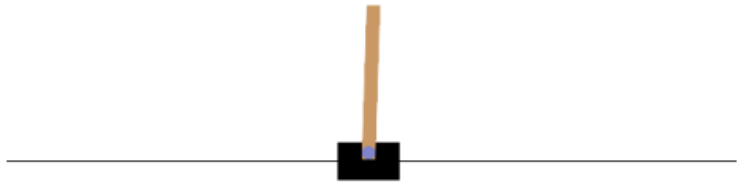


Learnt mixture

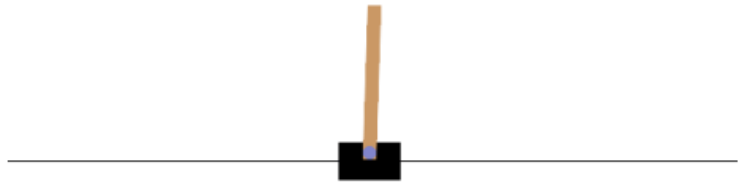
Simulation on Cartpole-v1 – Softmax PG

$K_{opt} + \Delta$

K_{opt}



Softmax PG



Learnt mixture

Base controllers

Thank You !

For more details: Visit our poster !