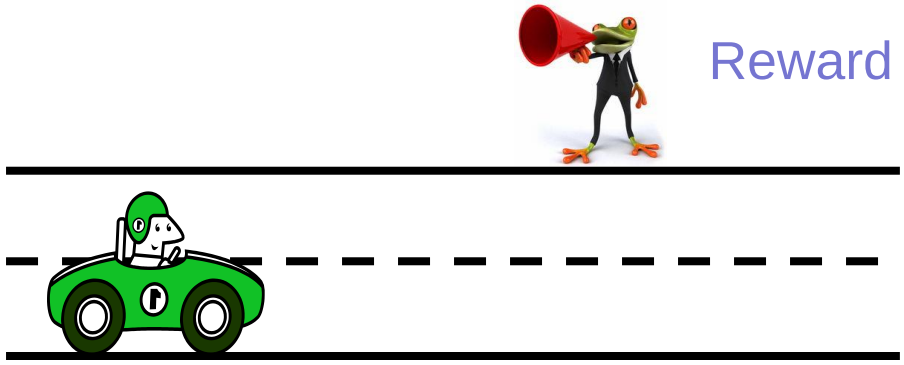
An Optimal Online Policy Gradient Algorithm for Continuous State and Action Markov Decision Processes

Yao Ma, Tingting Zhao, and Masashi Sugiyama
Department of Computer Science, Tokyo Institute of Technology

IBISML

Motivating Example

- Long road trip over some period of time T



State: Position
Action: Select road segment
Goal: Find the best strategy that minimizes the delay

Online Markov Decision Process

- State space
- Action space
- Transition
- Policy
- Reward function is time dependent

Our Method

- Online policy gradient

- Simple and natural
- Continuous states and actions
- Regret against the best offline policy is O(log T)