Online Convex Optimization in Adversarial MDPs
Aviv Rosenberg  Yishay Mansour

Motivation:
- MDPs are very popular but don’t consider time-changing environments
- BGP Routing is a great motivating example

Model:
- Episodic MDP
- Transition Function is fixed but unknown to the learner
- Sequence of loss functions is chosen by an adversary
- Success is measured by the regret – comparing to the best policy in hindsight
Problem Reformulation:
- The learner picks policies or occupancy measures equivalently
- Picking occupancy measures makes this an instance of online convex optimization

Algorithm:
- Basic idea: run online mirror descent
- Problem: unknow transition function means we don’t know if an occupancy measure is legal
- Solution: maintain confidence sets that contain the MDP with high probability

Occupancy measure is a probability distribution over the state-action pairs
Online Convex Optimization in Adversarial MDPs
Aviv Rosenberg       Yishay Mansour

Challenges:
- Efficient implementation of the algorithm
- Regret analysis

Contributions:
- Handling performance criteria that are convex with respect to the occupancy measures
- High confidence regret bound of \( O(H|S|\sqrt{|A|T}) \)

Performance criterion is a function that aggregates all the losses of a single episode.
Examples involve risk-sensitivity and robustness.

Previous state-of-the-art:
- Based on Follow the Perturbed Leader
- Regret bound of \( O(H|S||A|\sqrt{T}) \) in expectation