Can we plan efficiently in large MDPs with only weak linear function approximation and no restrictions on MDP dynamics?

Assumption: Core States

A small subset of states (of size \( m \)) whose features’ convex hull covers all other state features

Planning in Large MDPs

Avoid scaling with number of states, or exponential scaling in horizon (\( H = 1/(1 - \gamma) \) is the effective horizon)

\( \checkmark \) Impossible without additional assumptions!

Need \( (1/\varepsilon)^m \) samples for \( \varepsilon \)-suboptimal policy [Kearns, et al., 2002]

\( \times \) Impossible with weak function approximation, if policy must be \( \varepsilon \)-suboptimal [Du, et al., 2020]

\( \checkmark \) Possible for \( (1/\varepsilon)^{m(H/\sqrt{d})} \)-suboptimal policies, but requires value functions of all policies to be representable with low error [Lattimore, et al., 2020; Van Roy & Dong, 2019]

\( \checkmark \) Possible with strong assumptions on MDP dynamics (linear MDPs, low Bellman rank, etc.)

CoreStoMP

A Saddle-Point Algorithm for Planning with Core States

Based on Relaxed Approximate Linear Program [Lakshminarayanan, et al., 2018]

Uses Stochastic Mirror-Prox to approximately solve saddle-point formulation of problem

Gradient estimates come from simulator

Main Result

Running CoreStoMP on state \( s \) for \( T \) iterations:

- Uses the simulator \( O(mA T) \) times
- Outputs random action \( a \) with
  \[ \sum_{s \in S} \| v^*(s) - v_\pi(s) \| \leq O \left( \frac{\varepsilon_{\text{approx}}}{1 - \gamma} \right) + O \left( \frac{1}{(1 - \gamma)^2} \sqrt{\frac{m}{T}} \right) \]
- Results in policy \( \pi \) with value loss
  \[ \max_{s \in S} \| v^*(s) - v_\pi(s) \| \leq O \left( \frac{\varepsilon_{\text{approx}}}{1 - \gamma} \right) + O \left( \frac{1}{(1 - \gamma)^2} \sqrt{\frac{m}{T}} \right) \]

References