On Model Accuracy and Planning Horizon

Nan Jiang, Alex Kulesza, Satinder Singh
Computer Science & Engineering, University of Michigan
Richard Lewis
Department of Psychology, University of Michigan

Abstract

For MDPs with a large discount factor, it is common to use reduced horizon in planning to speed computation. However, when the model available to the agent is inaccurate, the policy found using a shorter planning horizon can be actually better than the one learned with the true horizon. The reason, we argue, is that planning horizon is a complexity control parameter for the policy class to be learned. We propose two complexity measures controlled by planning horizon. The implications provided by the loss bounds based on the proposed complexity measures are empirically verified.

Outline

Given MDP $M = \langle S, A, T, R, \gamma_{\text{eval}} \rangle$, consider an agent that

- plans with an inaccurate model $\widehat{M}$, estimated from dataset of size n,
- uses a discount factor $\gamma \leq \gamma_{\text{eval}}$ in planning,
- outputs $\pi^*_{\widehat{M},\gamma}$ (i.e. the **C**ertainty-**E**quivalent policy)

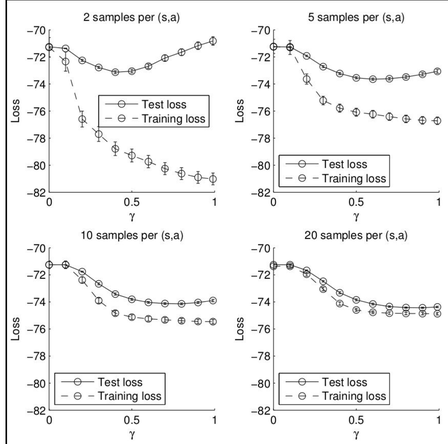
Question: what is the optimal γ that minimizes planning loss, namely

$$\gamma^* = \arg \min_{\gamma \in [0, \gamma_{\text{eval}}]} \left\| V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma_{\text{eval}}} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{\text{eval}}} \right\|_\infty$$

Our work:

1. Empirical observation: $\gamma^* < \gamma_{\text{eval}}$; $n \uparrow \Rightarrow \gamma^* \downarrow$
2. Theoretical explanation: γ is a complexity control parameter

Empirical Illustration



Test loss: $-V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{\text{eval}}}$

Training loss: $-V^{\pi^*_{\widehat{M},\gamma}}_{\widehat{M},\gamma_{\text{eval}}}$

Experiment details

- A synthetic MDP with 10 states, 2 actions
- Data: n samples per (s,a) for n = 2, 5, 10, 20
- $\widehat{M} = \langle S, A, \widehat{T}, R, \gamma \rangle$, where $\widehat{T}(s,a,s') = n(s,a,s') \,/\, n(s,a)$
- Average loss over 1000 draws of data is reported

Theoretical framework

Empirical risk minimization	Certainty-equivalent planning
• A dataset	• Empirical model $\widehat{M}$
• A class of hypotheses	• A class of policies
• Select the hypothesis that minimizes training error	• Select the policy that maximizes $V^\pi_{\widehat{M},\gamma}$

Loss bound → Approximation error ↓ − Estimation error ↑↓ → complexity of hypotheses class ↑ dataset size ↑

⇩

An intermediate complexity is optimal and it grows with dataset size

Loss bound → ? → complexity of policy class ? ⇐ γ ↑

⇩

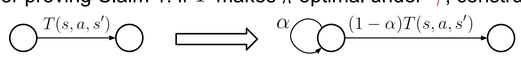
$\gamma^* < \gamma_{\text{eval}}$

$n \uparrow \Rightarrow \gamma^* \downarrow$

A Counting Complexity Measure

- Observation: fixing R and γ, some π can never be optimal
- Define $\Pi_{R,\gamma}$ as the set of policies that *can* be optimal under R and γ, i.e.
  $$\Pi_{R,\gamma} = \{\pi : \exists T, \text{ s.t. } \pi \text{ is optimal in } \langle \mathrm{S}, \mathrm{A}, \mathrm{T}, \mathrm{R}, \gamma \rangle\}$$
- Theorem 1 ($|\Pi_{R,\gamma}|$ increasing with γ)
  1. $\gamma \leq \gamma' \Rightarrow \Pi_{R,\gamma} \subseteq \Pi_{R,\gamma'}$
  2. $|\Pi_{R,0}| = 1$, *if* $\forall s \in S$, $\arg\max_{a \in A} R(s,a)$ *is unique.*
  3. $\exists \gamma$, s.t. $|\Pi_{R,\gamma}| \geq |A|^{|S|-2}$, *if* $\exists\, s, s' \in S$, $\max_{a \in A} R(s,a) > \max_{a' \in A} R(s',a')$.

Idea for proving Claim 1: if T makes π optimal under γ, construct T':



Any policy's value function w.r.t.(T,γ) is proportional to that w.r.t.(T',γ').

- Theorem 2 (Planning loss bound). w.h.p. planning loss is bounded by

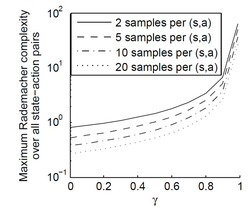
$\downarrow -$ ⇐ $\gamma \uparrow$ $n \uparrow$ ⇒ $\uparrow$

$$\underbrace{\frac{\gamma_{\text{eval}} - \gamma}{(1-\gamma_{\text{eval}})(1-\gamma)} R_{\max}}_{\max_{\pi:S\to A} \left\| V^\pi_{M,\gamma_{\text{eval}}} - V^\pi_{M,\gamma} \right\|_\infty} + \underbrace{\frac{2R_{\max}}{(1-\gamma)^2} \sqrt{\frac{1}{2n} \log \frac{2|S||A||\Pi_{R,\gamma}|}{\delta}}}_{\left\| V^{\pi^*_{M,\gamma}}_{M,\gamma} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma} \right\|_\infty}$$

Rademacher Complexity Measure

- More general: does not require fixed reward
- Define function class $\mathcal{F}_{M,\gamma} = \{f^\pi_{M,\gamma} : \pi \in S \to A\}$, where $f^\pi_{M,\gamma}(r,s') = r + \gamma V^\pi_{M,\gamma}(s')$
- Complexity measure: $\max_{s \in S, a \in A} \widehat{\mathfrak{R}}_{D_{s,a}}(\mathcal{F}_{M,\gamma})$
  - $D_{s,a}$: reward & next-state pairs sampled from (s,a)

$$\widehat{\mathfrak{R}}_{D_{s,a}}(\mathcal{F}_{M,\gamma}) = \mathbb{E}_{\substack{\sigma_i \sim Unif\{-1,1\} \\ i=1,2,\ldots,n}} \left\{ \sup_{\pi: S \to A} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i f^\pi_{M,\gamma}(r_i, s'_i) \right) \right\}$$



- Increasing with γ : shown empirically
- Planning loss bound:

$\downarrow -$ ⇐ $\gamma \uparrow$ $n \uparrow$ ⇒ $\uparrow$

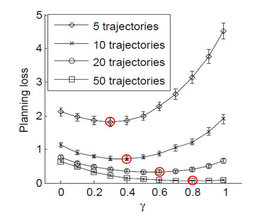
$$\frac{\gamma_{\text{eval}} - \gamma}{(1-\gamma_{\text{eval}})(1-\gamma)} R_{\max} + \frac{1}{1-\gamma} \left(2 \max_{s \in S, a \in A} \widehat{\mathfrak{R}}_{D_{s,a}}(\mathcal{F}_{M,\gamma}) + \frac{3R_{\max}}{1-\gamma} \sqrt{\frac{1}{2n} \log \frac{4|S||A|}{\delta}} \right)$$

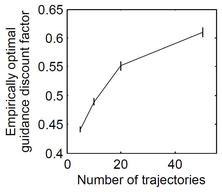
⋁|

$$\max_{s \in S, a \in A} \max_{\pi: S \to A} \left| \frac{1}{n} \sum_{(r,s') \in D_{s,a}} f^\pi_{M,\gamma}(r,s') - \mathbb{E}_{(r,s') \sim \mathbb{P}_{s,a}} \left\{ f^\pi_{M,\gamma}(r,s') \right\} \right|$$

Experiments in more realistic settings

- Synthetic MDPs, data collected via trajectories



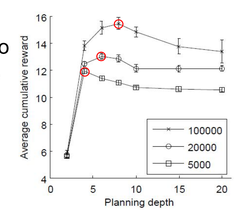


Left: planning loss vs γ for a single MDP.

Right: average γ^* over 1000 random MDPs.

Experiment details: MDPs with 10 states, 2 actions, 5 next-states, T and R generated randomly. Trajectories start from random initial state followed by purely random exploration.

- Optimal planning depth in UCT
  - UCT has a perfect simulator, but still needs to translate sample trajectories to action choice
  - #trajectories ↑ ⇒ optimal planning depth ↑



Experiment details: *RockSample* (7x8) domain, exploration parameter separately optimized for each (planning depth, #trajectories) pair.