The Dependence of Effective Planning Horizon on Model Accuracy

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis

University of Michigan
{nanjiang,kulesza,baveja,rickl}@umich.edu

ABSTRACT

For Markov decision processes with long horizons (i.e., discount factors close to one), it is common in practice to use reduced horizons during planning to speed computation. However, perhaps surprisingly, when the model available to the agent is estimated from data, as will be the case in most real-world problems, the policy found using a shorter planning horizon can actually be *better* than a policy learned with the true horizon. In this paper we provide a precise explanation for this phenomenon based on principles of learning theory. We show formally that the planning horizon is a complexity control parameter for the class of policies to be learned. In particular, it has an intuitive, monotonic relationship with a simple counting measure of complexity, and that a similar relationship can be observed empirically with a more general and data-dependent Rademacher complexity measure. Each complexity measure gives rise to a bound on the planning loss predicting that a planning horizon shorter than the true horizon can reduce overfitting and improve test performance, and we confirm these predictions empirically.

Categories and Subject Descriptors

I.2.8 [**Problem Solving, Control Methods, and Search**]: Dynamic programming

Keywords

Reinforcement learning; over-fitting; discount factor

1. INTRODUCTION

When planning with Markov decision processes (MDPs), we distinguish between two different horizons (or discount factors). The *evaluation horizon*, specified by the problem formulation, is part of the definition of the ultimate measure of performance for a policy and cannot be changed. The *planning horizon*, on the other hand, is a parameter supplied to the planning algorithm; it affects the resulting policy but need not match the evaluation horizon. Generally, the deeper or longer the planning horizon, the greater the computational expense of computing a policy [1, 2], while in principle the shallower or shorter the planning horizon (relative to the evaluation horizon), the more suboptimal the resulting policy is likely to be [1]. Thus, there is a tradeoff between computation and optimality that is relatively well-understood in cases where the model used for planning is accurate.

In this paper, we argue that there is another important reason to use shorter planning horizons in the more realistic case where the model used for planning is estimated from data: avoiding overfitting. Specifically, we show formally that the planning horizon controls the complexity of the policy class—shorter planning horizons define less complex policy classes. As in supervised learning, the optimal complexity (and therefore the optimal planning horizon) depends on the quantity of data used to estimate the model.

We explore two measures of complexity in this paper. The first is a simple and intuitive counting measure that we show is monotonically related to the planning horizon. The second is a Rademacher complexity measure [3], which affords a more general analysis. For each measure we prove a bound on the planning loss given a particular choice of planning horizon. Each bound has two terms that depend in opposite ways on the planning horizon: one prefers the longest possible planning horizon (up to the true horizon), encouraging fidelity to the ultimate evaluation metric, while the other encourages the shortest possible planning horizon, keeping the policy class simple and thereby reducing the possibility of overfitting. In general, the bounds suggest that some intermediate planning horizon will be optimal. We verify these predictions empirically, showing that even in the absence of computational constraints it can be beneficial to use a reduced planning horizon.

Section 2 provides background on planning in MDPs. Sections 3 and 4 formalize the counting complexity measure. Rademacher complexity is discussed in Section 5, and Section 6 provides experimental validation of our claims.

2. PRELIMINARIES: MDP PLANNING

An MDP specifies the agent-environment interaction model as a 5-tuple $M = \langle S, A, T, R, \gamma_{\text{eval}} \rangle$, where S is the state space, A is the action space, $T : S \times A \times S \to [0, 1]$ is the transition probability function, $R : S \times A \to \mathbb{R}$ is the expected reward function, and γ_{eval} is the evaluation discount factor. The agent's goal is to maximize expected utility, the expected value of the sum of future reward discounted by γ_{eval}. We assume rewards are bounded in the interval $[0, R_{\max}]$. A policy $\pi : S \to A$ is a mapping from states to actions. A policy that when followed maximizes expected utility in M is an optimal policy; we denote such a policy as $\pi^*_{M,\gamma_{\text{eval}}}$ to make explicit its dependence on γ_{eval}. We denote the value

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*

function of policy π evaluated in MDP M using discount factor γ as $V^{\pi}_{M,\gamma} \in \mathbb{R}^{|S|}$.

Certainty-equivalence control. In practical settings, we rarely know the true model of the agent-environment interaction. Here, we are interested in the case where the model is estimated from experience data in the real world; scarcity of data then implies that our model will only be approximate. In *certainty-equivalence control* we act according to the policy that is optimal with respect to the inaccurate model used for planning. Hereafter, we will be concerned with the performance of the certainty-equivalence policy derived from an estimated model $\widehat{M}$ using a *guidance discount factor* γ (which might not be equal to γ_{eval}). (If $\widehat{M} = M$ and $\gamma = \gamma_{\text{eval}}$, the certainty-equivalence policy is optimal.) In particular, we will consider $\widehat{M}$ that differs from M in T and R, and $\gamma \leq \gamma_{\text{eval}}$.

Evaluation. We emphasize that the certainty-equivalence policy computed using γ in model $\widehat{M}$ will nonetheless be evaluated in M using γ_{eval}. We capture this explicitly in our definition of the planning loss as the largest (over states) absolute difference in the values of the optimal policy $\pi^*_{M,\gamma_{\text{eval}}}$ and the CE-control policy $\pi^*_{\widehat{M},\gamma}$ when each is evaluated in the true environment M with the evaluation discount factor γ_{eval}. Formally, we have

$$\textbf{Planning loss}: \quad \|V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma_{\text{eval}}} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{\text{eval}}}\|_\infty \ , \qquad (1)$$

where $\|\cdot\|_\infty$ denotes the L_∞ norm of a vector, i.e., the largest absolute value of any entry.

Discount factors and planning horizon. When computing a policy with guidance discount factor γ, there is an implicit notion of planning horizon. The larger γ, the longer the planning horizon, because rewards further into the future have an effect on the choice of optimal action in the current state. Indeed, in tree-search based planning algorithms such as UCT [4, 2], γ is explicitly translated into a planning horizon (usually by setting it to $\frac{1}{(1-\gamma)}$). *Here, we use guidance discount factor and planning horizon interchangeably with the understanding that the actual use depends on the nature of the planning algorithm.*

Optimal guidance discount factor. The decoupling of γ_{eval} and γ is fundamental to our work. The former is specified by the MDP, while the latter is a parameter under the control of the planning algorithm. If $\widehat{M} = M$, the only reason for $\gamma < \gamma_{\text{eval}}$ would be to obtain computational savings (at the expense of acting suboptimally). Our aim is to show that when $\widehat{M} \neq M$ there is another important reason to pick $\gamma < \gamma_{\text{eval}}$.

Given M and $\widehat{M}$, an optimal guidance discount factor can be defined as follows:

$$\gamma^* = \operatorname*{arg\,min}_{0 \leq \gamma \leq \gamma_{\text{eval}}} \|V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma_{\text{eval}}} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{\text{eval}}}\|_\infty \ . \qquad (2)$$

This is the discount factor the certainty-equivalence planner should use to minimize planning loss. (In general, there will be a range of optimal values for γ^*; for computational reasons it is natural to pick the smallest value in that range.)

3. PLANNING HORIZON & COMPLEXITY

Equation 2 suggests that $\gamma^* < \gamma_{\text{eval}}$ might be optimal—and indeed this is often observed in practice—but we do not yet have a clear intuition about when or why that would be true. We offer the following explanation: γ is a complexity control parameter for the policy class. Specifically, we will show in this section that γ monotonically controls the number of policies that can be optimal given a fixed state space, action space, and reward function. When $\widehat{M}$ is estimated from a limited data set, we can therefore avoid overfitting in policy selection by restricting the number of available policies through γ. (In Section 5, we will extend this intuition to a more sophisticated Rademacher measure.)

In the traditional empirical risk minimization setting for supervised learning, training data are used to evaluate the models in a given model class, and the model with the lowest training error is selected [5]. Overfitting occurs when the model class is too complex compared to the effective size of the dataset, and one way to avoid overfitting is to limit the complexity of the model class.

We draw analogies to four elements in this scenario: (1) the size of the dataset, (2) the complexity of the model class, (3) empirical risk minimization as a method for selecting a model from the class of models, and (4) some way to control model complexity. In our planning setting, the size of the dataset corresponds to the number of samples used to estimate $\widehat{M}$. We assume that for every state-action pair (s, a), we observe n samples of the successor state drawn from the true transition function. (For now, we assume that the rewards R are known exactly.) The model class in our setting is the set of policies that might be optimal in $\widehat{M}$, and, initially, the complexity of the model class corresponds to the *number* of policies being searched over. Empirical risk minimization corresponds to selecting the optimal policy for $\widehat{M}$, as achieved by certainty-equivalence planning. These three correspondences are evident. It remains to show that reducing the guidance discount factor γ corresponds to reducing the size of the policy class being searched over by planning. Theorem 1 shows that this is indeed the case.

Theorem 1. *For any fixed state space S, action space A, and reward function R, define*

$$\Pi_{R,\gamma} = \{\pi : \exists\ T \ \textit{s.t.}\ \pi \ \textit{is optimal in}\ \langle S, A, T, R, \gamma\rangle\}. \quad (3)$$

Then the following claims hold:

1. $\left|\Pi_{R,0}\right| = 1$
   if, for all $s \in S$, $\arg\max_{a \in A} R(s, a)$ is unique.
2. $\forall \gamma, \gamma' : 0 \leq \gamma \leq \gamma' < 1, \quad \Pi_{R,\gamma} \subseteq \Pi_{R,\gamma'}$.
3. $\exists\ \gamma \in [0, 1), \left|\Pi_{R,\gamma}\right| \geq |A|^{|S|-2}$
   if $\exists\ s, s' \in S, \ \max_{a \in A} R(s, a) > \max_{a' \in A} R(s', a')$.

The assumption for claim 1 ensures that there are no ties in the maximal reward for each state, and the assumption for claim 3 requires that one cannot obtain the maximal reward at every state. Note that $\Pi_{R,\gamma}$ counts policies that are optimal as T is allowed to vary arbitrarily, but explicitly depends on the fixed, known reward function R. (If R were allowed to vary with T, then every policy could be optimal at every γ.) In Section 5 we will show how this restriction can be lifted.

Taken together, the three claims of Theorem 1 show that γ monotonically adjusts the size of the policy class from 1 to at least $|A|^{|S|-2}$, which is "almost all" of the $|A|^{|S|}$ possible policies. Thus the choice of guidance discount factor tightly

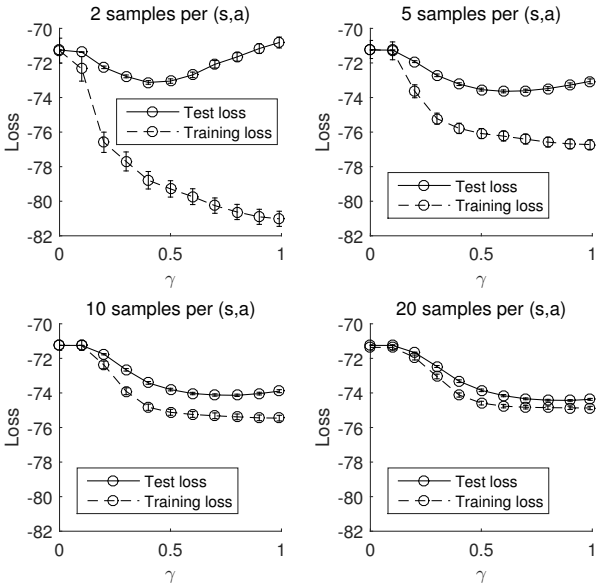

Figure 1: Learning curves as a function of γ, the guidance discount factor. For each MDP M sampled from the RANDOM-MDP distribution specified in Section 6 we build $\widehat{M}$ by sampling each state-action pair $n = 2$, 5, 10, or 20 times; the different subgraphs correspond to different values of n. The reward function is assumed known, and $\gamma_{\text{eval}} = 0.99$. The training loss is the negative value of the certainty-equivalence policy on the estimated model $\widehat{M}$: $-\frac{1}{|S|}\sum_{s\in S} V^{\pi^*_{\widehat{M},\gamma}}_{\widehat{M},\gamma_{\text{eval}}}(s)$, and the test loss is the negative value of that same policy on the actual MDP M: $-\frac{1}{|S|}\sum_{s\in S} V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{\text{eval}}}(s)$. Confidence intervals are computed using 1000 i.i.d. draws of M.

controls complexity. Figure 1 illustrates this by showing that, as γ varies from 0 to γ_{eval}, we recover the traditional learning curves from supervised learning (see caption for details). Training loss decreases monotonically as γ increases, while test loss is U-shaped, indicating that an overly large γ causes overfitting. We can also see in Figure 1 that the location of the minimum of the test loss curve—that is, the optimal test γ—shifts to the right as we get more data.

We now prove the three claims in turn. The first is straightforward; given the stated assumption, the optimal policy does not depend on T when $\gamma = 0$. Thus, the policy that picks the action with the highest immediate reward is the only one that can be optimal.

PROOF OF THEOREM 1, CLAIM 2. We will prove that for $\gamma \le \gamma'$, $\pi \in \Pi_{R,\gamma} \Rightarrow \pi \in \Pi_{R,\gamma'}$. Let T be a transition function for which π is optimal in $\langle S, A, T, R, \gamma\rangle$. We will construct T' such that the MDP $M' = \langle S, A, T', R, \gamma'\rangle$ has the property that for all $\pi' : S \to A$,

$$V^{\pi'}_{M',\gamma'} = cV^{\pi'}_{M,\gamma} \ , \tag{4}$$

where c is a positive constant that only depends on γ and γ'. Consequently, π is also optimal in M'. Let $T'(s,a,s') = (1-\alpha)T(s,a,s') + \alpha\mathbb{I}(s = s')$, where $\mathbb{I}(\cdot)$ is the indicator function and α is a scalar in the range $[0,1]$. That is, T' is a transition function where, with probability $1-\alpha$, transitions behave according to T, but with probability α, a state simply transitions to itself. Recall that

$$V^{\pi'}_{M,\gamma} = (I - \gamma[T^{\pi'}])^{-1}R^{\pi'} \tag{5}$$

$$V^{\pi'}_{M',\gamma'} = (I - \gamma'[T'^{\pi'}])^{-1}R^{\pi'} \ , \tag{6}$$

where $[T^{\pi'}]$ is the $|S|\times|S|$ matrix with $[T^{\pi'}](s,s') = T(s,\pi'(s),s')$ and $R^{\pi'}$ is the $|S|\times 1$ vector with $R^{\pi'}(s) = R(s,\pi'(s))$. We have

$$[T'^{\pi'}] = (1-\alpha)[T^{\pi'}] + \alpha I \ , \tag{7}$$

hence

$$\begin{aligned} V^{\pi'}_{M',\gamma'} &= \left(I - \gamma'\big((1-\alpha)[T^{\pi'}] + \alpha I\big)\right)^{-1} R^{\pi'} \\ &= \left((1-\gamma'\alpha)I - \gamma'(1-\alpha)[T^{\pi'}]\right)^{-1} R^{\pi'} \\ &= \frac{1}{1-\gamma'\alpha}\left(I - \frac{\gamma'(1-\alpha)}{1-\gamma'\alpha}[T^{\pi'}]\right)^{-1} R^{\pi'} \ . \end{aligned}$$

Letting $\frac{\gamma'(1-\alpha)}{1-\gamma'\alpha} = \gamma$, we get $\alpha = \frac{1-\gamma/\gamma'}{1-\gamma}$, which is between 0 and 1 since $0 \le \gamma \le \gamma' < 1$, and thus

$$V^{\pi'}_{M',\gamma'} = \frac{1-\gamma}{1-\gamma'}V^{\pi'}_{M,\gamma} \ . \tag{8}$$

This completes the proof. □

PROOF OF THEOREM 1, CLAIM 3. The proof is by construction. Let (s^*, a^*) be a state-action pair that achieves the highest reward among all state-action pairs. Let s' be a state whose maximal reward action a' gives reward strictly less than $R(s^*, a^*)$. Such a state always exists under the assumption for this claim in the theorem. Consider an arbitrary policy π, with the only constraints that $\pi(s^*) = a^*$ and $\pi(s') = a'$. Then the following transition function makes π optimal for large enough γ:

$$\forall\ s \in S \quad T(s,a,\cdot) = \begin{cases} \mathbf{1}_{s^*} & \text{if } a = \pi(s), s \ne s' \\ \mathbf{1}_{s'} & \text{otherwise} \end{cases} \tag{9}$$

where $\mathbf{1}_{(\cdot)}$ denotes the delta distribution. The optimality of π at s^* and s' is trivial, as both states are absorbing and π chooses the action that maximizes immediate reward. In any other state s, we show that π is optimal by comparing the optimal Q-value of $(s,\pi(s))$ to that of (s,a) for any other action a:

$$Q^*(s,\pi(s)) = R(s,\pi(s)) + \frac{\gamma}{1-\gamma}R(s^*,a^*), \tag{10}$$

$$Q^*(s,a) = R(s,a) + \frac{\gamma}{1-\gamma}R(s',a'). \tag{11}$$

We know $R(s^*,a^*) - R(s',a') > 0$, and as γ approaches one, $\gamma/(1-\gamma)$ tends to infinity, so for sufficiently large γ we can guarantee that $Q^*(s,\pi(s)) > Q^*(s,a)$. Recall that we constrained π in only two states, hence the number of such policies is $|A|^{|S|-2}$. □

4. PLANNING LOSS BOUND

Completing the connection to model class complexity in supervised learning, we show that the loss of the certainty-equivalence policy for $\widehat{M}$ is bounded, with high probability, in terms of the policy class complexity $|\Pi_{R,\gamma}|$. This is analogous to a standard generalization bound [6], and implies that an intermediate value of γ will generally be optimal; moreover, as the amount of data (n) increases, so does the optimal γ.

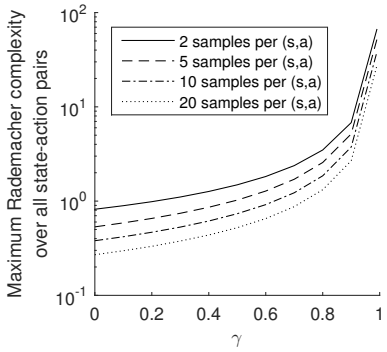

Figure 2: The relationship between empirical Rademacher complexity ($\max_{s,a} \widehat{\mathfrak{R}}_{D_{s,a}}(\mathcal{F}_{M,\gamma})$)[1]and guidance discount factor γ. Results are averaged over 10,000 MDPs sampled from the RANDOM-MDP distribution (see Section 6), and both the reward and transition functions are estimated from samples.

Theorem 2. *Let M be an MDP with non-negative rewards and evaluation discount factor γ_{eval}. Let $\widehat{M}$ be an MDP comprising the true reward function of M and a transition function estimated from n samples for each state-action pair. Then certainty-equivalence planning with $\widehat{M}$ using guidance discount factor $\gamma \leq \gamma_{eval}$ has planning loss*

$$\|V^{\pi^*_{M,\gamma_{eval}}}_{M,\gamma_{eval}} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{eval}}\|_\infty \leq \frac{\gamma_{eval}-\gamma}{(1-\gamma_{eval})(1-\gamma)} R_{\max} + \frac{2R_{\max}}{(1-\gamma)^2}\sqrt{\frac{1}{2n}\log\frac{2|S||A||\Pi_{R,\gamma}|}{\delta}} \quad (12)$$

with probability at least $1-\delta$.

The proof of the theorem is in Appendix A. The upper bound in Theorem 2 has two terms. The first is a bound on the planning loss incurred by using the guidance discount factor γ instead of the evaluation discount factor γ_{eval} in the true M. This term goes to zero as γ increases and approaches γ_{eval}. The second term isolates the planning loss due to the use of $\widehat{M}$ instead of M, but does not depend on γ_{eval}. In contrast to the first term, this term increases with γ, since greater policy class complexity allows performance on M and $\widehat{M}$ to diverge more dramatically. The dependence on the policy complexity $\left|\Pi_{R,\gamma}\right|$ is the novelty of our bound, compared to related work bounding loss by model errors or Bellman residuals [7, 8, 9].

The two terms in the bound of Theorem 2 depend in opposite ways on γ, therefore the bound will be optimized at some intermediate value. As the amount of data n increases, the second term will shrink and the bound will prefer larger values of γ. We will observe this behavior empirically in Section 6.

5. RADEMACHER COMPLEXITY BOUND

In the previous sections we showed how $|\Pi_{R,\gamma}|$ can be used to bound the loss of certainty-equivalence planning. While this simple complexity measure has the advantage of being easy to interpret and allowed us to prove a clean, monotonic relationship with the guidance discount factor, the analysis required assuming that the reward function was known. Furthermore, hypothesis-counting measures of complexity are typically weak, whereas modern data-dependent measures can be significantly tighter and more sensitive [12].

In this section, we present an alternative analysis using a Rademacher complexity measure [3] that does not assume the reward function is known. We provide a loss bound parallel to that in Theorem 2, which is also optimized at an intermediate γ that increases with sample size.

Theorem 3. *Let M be an MDP with non-negative rewards and evaluation discount factor γ_{eval}. Let $\widehat{M}$ be an MDP comprising reward and transition functions estimated from n samples for each state-action pair. Then certainty-equivalence planning with $\widehat{M}$ using guidance discount factor $\gamma \leq \gamma_{eval}$ has planning loss*

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

with probability at least $1-\delta$, where

- $\mathcal{F}_{M,\gamma} = \{f^\pi_{M,\gamma} : \pi \in S \to A\}$, *with*
$$f^\pi_{M,\gamma}(r,s') = r + \gamma V^\pi_{M,\gamma}(s') \ .$$
- $D_{s,a}$ *is the set of n pairs of immediate reward & next-state sampled from (s,a) in dataset D.*
- $\widehat{\mathfrak{R}}_{D_{s,a}}(\mathcal{F}_{M,\gamma})$ *is the empirical Rademacher complexity of function class $\mathcal{F}_{M,\gamma}$ w.r.t. input points $D_{s,a}$, i.e.,*
$$\mathop{\mathbb{E}}_{\substack{\sigma_i \overset{i.i.d.}{\sim} \text{unif}\{-1,1\}\\ i=1,\ldots,n}} \left\{\sup_{f\in\mathcal{F}_{M,\gamma}} \frac{1}{n}\sum_{(r,s')\in D_{s,a}} \sigma_i f(r,s')\right\}. \quad (13)$$

The proof of the theorem is in Appendix B. The bound has the same decomposition as Theorem 2, but replaces the second term (loss due to planning with $\widehat{M}$ under γ) with a bound in terms of the Rademacher complexity of a function class $\mathcal{F}_{M,\gamma}$ in which each function corresponds to a policy in the MDP. For each state-action pair, the empirical model $\widehat{M}$ can be viewed as implicitly learning the expected values of all the functions in $\mathcal{F}_{M,\gamma}$ simultaneously from input samples $D_{s,a}$. The maximal deviation (over all functions) can be bounded by a state-action specific Rademacher complexity, and the worst case complexity (over all state-action pairs) translates to planning loss.

To show that the bound is optimized by an intermediate γ which increases with sample size n, it suffices to show that the second term increases with γ and decreases with n. This would be straightforwardly true if we knew that $\max_{s\in S, a\in A} \widehat{\mathfrak{R}}_{D_{s,a}}(\mathcal{F}_{M,\gamma})$ increased monotonically with γ in the manner of Theorem 1. We leave a formal result to future work; here we show empirically that the data-dependent Rademacher complexity is strongly and positively correlated with γ in practice: see Figure 2, where the relationship appears clearly monotonic. Thus Theorem 3 has the same qualitative interpretation as Theorem 2 while employing the more sensitive Rademacher measure.

[1]Since we cannot feasibly enumerate all possible values of $\{\sigma_i\}_{i=1}^n$ to compute the expectation in Equation 13, we take the standard approach and sample them uniformly to obtain an approximation [10, 11]. We found that 100 samples was sufficient to give low variance.

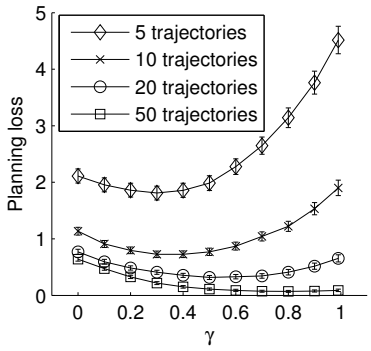

Figure 3: Planning loss as a function of γ for a single MDP drawn from RANDOM-MDP. From top to bottom, the curves correspond to increasing dataset sizes and are labeled by the number of trajectories in the dataset.

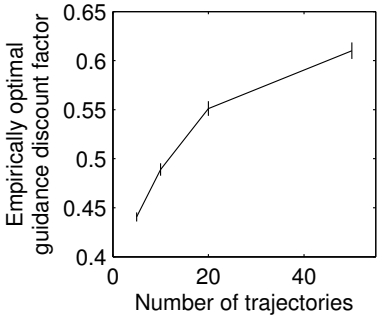

Figure 4: Optimal guidance discount factor as a function of dataset size, averaged over 1,000 MDPs from RANDOM-MDP and 1,000 datasets for each MDP. Higher values (closer to one) are optimal for minimizing the planning-loss of certainty-equivalence policies as the amount of data increases.

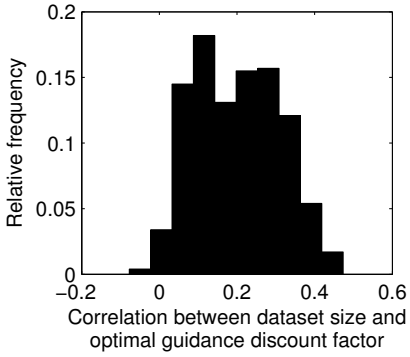

Figure 5: Histogram of the correlation between dataset size and γ^* over 1,000 randomly generated MDPs from RANDOM-MDP. For almost all the MDPs, there is a positive correlation between dataset size and γ^*, indicating that γ^* increasing with dataset size does not only hold in the average sense, but also applies to individual problems.

6. EXPERIMENTAL RESULTS

We now show experimentally that the phenomena predicted by the preceding theoretical discussion do, in fact, appear in practice. In particular, we will see that the optimal choice of guidance discount factor can be smaller than γ_{eval}, and as we increase the amount of data used to estimate the model, a larger γ tends to be preferable.

For these experiments we randomly sampled 1,000 MDPs with 10 states and 2 actions from a distribution we refer to as RANDOM-MDP, defined as follows. For each state-action pair (s, a), the distribution over the next state, $P(s, a, \cdot)$, is determined by choosing 5 non-zero entries uniformly from all 10 states, filling these 5 entries with values uniformly drawn from $[0, 1]$, and finally normalizing $P(s, a, \cdot)$. The mean rewards were likewise sampled uniformly and independently from $[0, 1]$, and the actual reward signals have additive Gaussian noise with standard deviation 0.1. For all MDPs we fixed $\gamma_{\text{eval}} = 0.99$.

For each generated MDP M, and for each value of $n \in \{5, 10, 20, 50\}$, we independently generated 1,000 data sets, each consisting of n trajectories of length 10 starting at uniformly random initial states and choosing uniformly random actions. While our theoretical results assume the data set comprises n samples for each state-action pair, for our experiments we chose to generate trajectories since for most applications they are a more realistic way to collect data. (We also performed the same experiments using samples of state-action pairs and the results were qualitatively similar.)

For each dataset D, we set $\widehat{M}$ to be the maximum-likelihood model; that is, the estimated reward $\widehat{R}(s, a)$ is the mean of the rewards observed at (s, a), and the estimated transition probability $\widehat{T}(s, a, s')$ is the number of times we observe the transition $(s, a) \to s'$ in D divided by the number of times we observe (s, a). If some (s, a) has never been seen in a dataset, we set $\widehat{R}(s, a) = 0.5$ and $\widehat{T}(s, a, s') = 1/|S|$.

For each value of $\gamma \in \{0, 0.1, 0.2, \ldots, 0.9, 0.99\}$, we compute the empirical loss

$$\frac{1}{|S|} \sum_{s \in S} \left(V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma_{\text{eval}}}(s) - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{\text{eval}}}(s) \right) , \qquad (14)$$

and pick the γ that minimizes the loss as an estimate of γ^* (see Equation 2), breaking ties randomly.

Figure 3 shows the empirical loss averaged over datasets as a function of the guidance discount factor γ for a characteristic MDP. Each curve in the figure corresponds to a particular number of trajectories as data. The error bars in this figure and elsewhere show 95% confidence intervals. We can see that the curves exhibit the U-shape predicted by the theory, with minimum planning loss achieved at some γ^* less than γ_{eval}. As expected, increasing dataset size reduces planning loss in general, and shifts γ^* to the right.

Figure 4 explicitly measures this shift by averaging the estimated γ^* across all 1,000 generated MDPs and their datasets. We can see clearly that as the amount of data increases, the optimal guidance discount factor increases as well. In the limit, of course, γ^* should equal γ_{eval}. However, for these values of dataset size the average γ^* is always significantly less than γ_{eval}; this means that using the true evaluation horizon for planning will lead to an increase in loss. While, conventionally, the use of a shorter horizon for planning has been justified based on computational savings, our result shows that in this setting it can decrease loss as well.

To complement the average-case analysis in Figure 4, Figure 5 shows the distribution of the correlation between

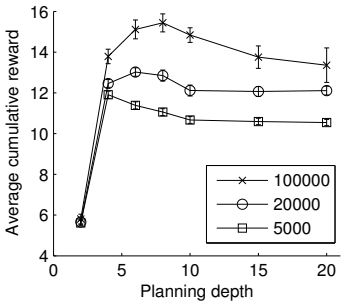

Figure 6: Performance of UCT as a function of planning depth. For each curve, the number of UCT trajectories is fixed to 5,000, 20,000, or 100,000. For each point on the graph, the UCB scalar has been separately optimized by sweeping through the values in $10 \cdot \exp\{-2, -1, 0, 1, 2\}$. For the 5,000 and 20,000 trajectory curves, each point is an average of 5,000 independent trials; the 100,000 trajectory curve is an average over 1,000 trials.

dataset size and γ^* over 1,000 individual MDPs. This correlation is positive with very high probability, implying that in almost all cases (under RANDOM-MDP) the theoretical relationship between dataset size and γ^* is borne out in practice.

6.1 Optimal planning depth in UCT

The previous experiments used small-state problems for which we could and did use perfect planning algorithms (value iteration) on the MDPs estimated from data. However, another common planning setting is one where we have an accurate (generative or probability) model, but the size of state space is so large that exact planning is impossible. Instead, incremental planning algorithms such as UCT must be used [2]. These algorithms repeatedly sample a search tree (rooted at the current state) that implicitly defines an inaccurate local model $\widehat{M}$ from which a policy is derived. Here we show that the main intuition obtained above—that planning horizon controls complexity, hence the more inaccurate the model the shorter the planning horizon that should be used—holds for UCT as well (see [13] for an alternative approach to controlling complexity in UCT via state abstractions).

In this setting, we do not have "data" in the sense of recorded experiences; instead, the accuracy of the local model is mediated by the number of trajectories sampled at the current state. Similarly, rather than manipulating a continuous discount factor γ we will control complexity via the planning depth, a discrete hyperparameter that sets the maximum length of the sampled trajectories. Our aim is to show that the relationship we have established between dataset size and discount factor for value iteration holds analogously between the number and depth of UCT trajectories.

We used a benchmark POMDP domain *RockSample* [14] and evaluated UCT's performance with different numbers of trajectories and different maximum depths. A detailed description of this infinite-sized belief-state space domain can be found in [15]; we used a map of size 7×8. Since this problem is episodic, we use the average cumulative reward per episode as our evaluation metric in place of planning loss (and so higher is better). Since episodes are usually on the order of hundreds of time steps, setting the planning depth to this level is computationally infeasible. However, Figure 6 shows that choosing a small planning depth not only speeds computation but also helps performance when the number of trajectories is limited.

In particular, an intermediate value of planning depth always achieves the highest cumulative reward. Moreover, as the number of trajectories grows from 5000 to 20000 to 100000, that optimal planning depth increases. This is qualitatively the same behavior we have seen before.

6.2 Selecting γ via cross-validation

We have seen that choosing $\gamma < \gamma_{\text{eval}}$ often improves performance, but how should we go about selecting the optimal γ in practice? In supervised learning, k-fold cross-validation is one of the most common techniques for selecting hyperparameters to avoid overfitting, and it is easy to apply here as well. (Indeed, we suspect cross-validation is often used in practice for choosing discount factors though we are unaware of any specific reference.)

Specifically, given a dataset D drawn from MDP M, we can split the sample trajectories into (state, action, reward, next-state) tuples, and then divide the tuples randomly into k folds of equal size, $D_1, \ldots, D_k$. For each fold $j = 1, 2, \ldots, k$, the validation model $\widehat{M}_j$ is defined to be the maximum-likelihood model learned from D_j, and the training model $\widehat{M}_{-j}$ is the one learned from $D \setminus D_j$. Then for each candidate γ, the validation value on fold j is given by

$$\text{ValidationValue}_j(\gamma) = \frac{1}{|S|} \sum_{s \in S} V^{\pi^*_{\widehat{M}_{-j},\gamma}}_{\widehat{M}_j,\gamma_{\text{eval}}}(s) \ . \qquad (15)$$

Cross-validation selects the value of γ that maximizes the validation value averaged over all folds.

However, there is a potential problem. While cross-validation produces unbiased estimates of loss in most supervised settings, in certainty-equivalence planning the use of a finite validation set biases our estimate of a policy's true value. This happens because, although the transition and reward functions in the validation model are themselves unbiased, the validation value of a policy is computed via a nonlinear matrix inverse (see Equation 6). Thus, for instance, a myopic policy may perform well in a model estimated from a small validation set due to reduced stochasticity. Under mild assumptions the bias can be shown to decrease much faster than variance when sample size is sufficiently large [16]; however, in practice our data sets are often relatively small.

Despite this caveat, our experiments in this section show that, at least in some instances, cross-validation can still be an effective practical tool for choosing γ. We leave the design and analysis of other cross-validation schemes for MDPs to future work; see also [17] for some discussion of this issue.

Figure 7 shows the average loss when choosing γ via 3-fold cross-validation compared to the losses obtained using fixed values of γ. We can see that small values of γ incur relatively large loss when there are sufficient samples, and large values of γ incur relatively large loss when there are few samples. In other words, no fixed γ dominates the others over all sample sizes. In contrast, cross-validation is able to achieve loss close to the best fixed γ at each sample size simultaneously by selecting γ adaptively as sample size changes.

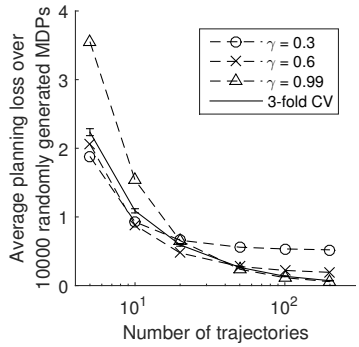

Figure 7: 3-fold cross-validation vs. fixed γ . Domain distribution and candidate guidance discount factors are the same as in Figure 4. We plot average loss as a function of sample size in terms of the number of trajectories.

7. RELATED WORK

The loss induced by a finite planning horizon is known as truncation loss (see related bounds in [1]). Separately, it is also well-understood how planning loss relates to model inaccuracy, which can come from estimation error when the model is constructed from data [7, 16], and/or approximation error when approximations are employed in planning (i.e., state abstractions [18]). It has been noted that such loss can have significant dependence on horizon [8, 19, 9]. To our knowledge, [20] is the first work to show how a short horizon can reduce loss when the model is inaccurate due to approximation errors. Our work explores a similar phenomenon due to estimation errors, and our analysis exploits the structure of these errors as well as established principles in supervised learning to obtain stronger claims about γ^* and dataset size.

8. CONCLUSION

We presented a connection between model complexity and planning horizon by developing a theoretical and empirical analogy to overfitting in supervised learning. We showed that the planning horizon controls the complexity of the policy space, and proved bounds on the loss of the certainty-equivalence policy using a simple counting complexity measure as well as Rademacher complexity. Each bound sets up a tradeoff between a term in which a larger planning horizon reduces the loss incurred in an accurate model and a term in which a smaller planning horizon reduces the complexity of the policy space and thereby controls overfitting. Empirical results confirm that the optimal choice of guidance discount factor is usually smaller than the discount factor defined by the problem, and that the optimal guidance discount factor increases with the amount of data.

Acknowledgement

This work was supported by NSF grant IIS 1319365. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the sponsors.

APPENDIX

A. PROOF OF THEOREM 2

We begin by proving Lemma 1 and Lemma 2.

Lemma 1. *For any MDP M with rewards in $[0, R_{\max}]$, $\forall \pi : S \to A$ and $\gamma \leq \gamma_{eval}$,*

$$V^{\pi}_{M,\gamma} \;\leq\; V^{\pi}_{M,\gamma_{eval}} \;\leq\; V^{\pi}_{M,\gamma} + \frac{\gamma_{eval} - \gamma}{(1-\gamma_{eval})(1-\gamma)} R_{\max} \;. \quad (16)$$

PROOF. The lower bound on $V^{\pi}_{M,\gamma_{\text{eval}}}$ follows directly from the assumption that reward is non-negative and that $\gamma \leq \gamma_{\text{eval}}$. For the upper bound, (letting $[T^{\pi}]$ denote the transition probability matrix for policy π)

$$\begin{aligned}
&\left\| V^{\pi}_{M,\gamma_{\text{eval}}} - V^{\pi}_{M,\gamma} \right\|_{\infty} \\
&= \left\| \sum_{t=1}^{\infty} ({\gamma_{\text{eval}}}^{t-1} - \gamma^{t-1})[T^{\pi}]^{t-1} R^{\pi} \right\|_{\infty} \\
&\leq \sum_{t=1}^{\infty} ({\gamma_{\text{eval}}}^{t-1} - \gamma^{t-1}) R_{\max} \\
&= \left(\frac{1}{1-\gamma_{\text{eval}}} - \frac{1}{1-\gamma}\right) R_{\max} = \frac{\gamma_{\text{eval}} - \gamma}{(1-\gamma_{\text{eval}})(1-\gamma)} R_{\max}. \square
\end{aligned}$$

Lemma 2. *Given true MDP M, let $\widehat{M}$ be an MDP comprising reward function $\widehat{R} = R$ and transition function $\widehat{T}$ estimated from n samples for each state-action pair, then*

$$\left\| V^{\pi^*_{M,\gamma}}_{M,\gamma} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma} \right\|_{\infty} \leq \frac{2R_{\max}}{(1-\gamma)^2} \sqrt{\frac{1}{2n} \log \frac{2|S||A||\Pi_{R,\gamma}|}{\delta}}$$

with probability at least $1 - \delta$.

We prove Lemma 2 with two additional lemmas: Lemma 3 translates planning loss to value error, and Lemma 4 relates value error to a Bellman-residual-like quantity that has a uniform deviation bound which depends on $\left|\Pi_{R,\gamma}\right|$.

Lemma 3. *For any $\widehat{M} = \langle S, A, \widehat{T}, \widehat{R}, \gamma \rangle$ with $\widehat{R}$ bounded by $[0, R_{\max}]$,*

$$\left\| V^{\pi^*_{M,\gamma}}_{M,\gamma} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma} \right\|_{\infty} \leq 2 \max_{\pi : S \to A} \left\| V^{\pi}_{M,\gamma} - V^{\pi}_{\widehat{M},\gamma} \right\|_{\infty} . \quad (17)$$

In particular, if $\widehat{R} = R$, we have

$$\left\| V^{\pi^*_{M,\gamma}}_{M,\gamma} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma} \right\|_{\infty} \leq 2 \max_{\pi \in \Pi_{R,\gamma}} \left\| V^{\pi}_{M,\gamma} - V^{\pi}_{\widehat{M},\gamma} \right\|_{\infty} . \quad (18)$$

PROOF. $\forall s \in S$,

$$\begin{aligned}
&V^{\pi^*_{M,\gamma},\gamma}_{M,\gamma}(s) - V^{\pi^*_{\widehat{M},\gamma},\gamma}_{M,\gamma}(s) \\
&= \left(V^{\pi^*_{M,\gamma},\gamma}_{M,\gamma}(s) - V^{\pi^*_{M,\gamma},\gamma}_{\widehat{M},\gamma}(s)\right) - \left(V^{\pi^*_{\widehat{M},\gamma},\gamma}_{M,\gamma}(s) - V^{\pi^*_{\widehat{M},\gamma},\gamma}_{\widehat{M},\gamma}(s)\right) + \\
&\quad \left(V^{\pi^*_{M,\gamma},\gamma}_{\widehat{M},\gamma}(s) - V^{\pi^*_{\widehat{M},\gamma},\gamma}_{\widehat{M},\gamma}(s)\right) \\
&\leq \left(V^{\pi^*_{M,\gamma},\gamma}_{M,\gamma}(s) - V^{\pi^*_{M,\gamma},\gamma}_{\widehat{M},\gamma}(s)\right) - \left(V^{\pi^*_{\widehat{M},\gamma},\gamma}_{M,\gamma}(s) - V^{\pi^*_{\widehat{M},\gamma},\gamma}_{\widehat{M},\gamma}(s)\right) \\
&\leq 2 \max_{\pi \in \left\{\pi^*_{\widehat{M},\gamma}, \pi^*_{M,\gamma}\right\}} \left| V^{\pi}_{M,\gamma}(s) - V^{\pi}_{\widehat{M},\gamma}(s) \right| .
\end{aligned}$$

Equation 17 follows from taking max over all states on both sides of the inequality and noticing that the set of all policies is a trivial superset of $\left\{\pi^*_{\widehat{M},\gamma}, \pi^*_{M,\gamma}\right\}$. If $\widehat{R} = R$, the bound can be tightened since $\left\{\pi^*_{\widehat{M},\gamma}, \pi^*_{M,\gamma}\right\} \subseteq \Pi_{R,\gamma}$ and Equation 18 follows. $\square$

Lemma 4. *For any $\widehat{M} = \langle S, A, \widehat{T}, \widehat{R}, \gamma\rangle$ with $\widehat{R}$ bounded by $[0, R_{\max}]$, $\forall \pi : S \to A$,*

$$\left\|Q^{\pi}_{M,\gamma} - Q^{\pi}_{\widehat{M},\gamma}\right\|_\infty \leq \frac{1}{1-\gamma} \max_{s\in S, a\in A} \left|\widehat{R}(s,a) + \gamma\langle \widehat{T}(s,a,\cdot), V^{\pi}_{M,\gamma}\rangle - Q^{\pi}_{M,\gamma}(s,a)\right|.$$

PROOF. Given any policy π, define state-action value functions $Q_0, Q_1, Q_2, \ldots, Q_m, \ldots$ such that $Q_0 = Q^{\pi}_{M,\gamma}$, and

$$Q_m(s,a) = \widehat{R}(s,a) + \gamma\langle \widehat{T}(s,a,\cdot), V_{m-1}\rangle,$$

where $V_{m-1}(s) = Q_{m-1}(s, \pi(s))$. Notice that

$$\begin{aligned}
&\|Q_m - Q_{m-1}\|_\infty \\
&= \gamma \max_{s\in S, a\in A} \left|\langle \widehat{T}(s,a,\cdot), (V_{m-1} - V_{m-2})\rangle\right| \\
&\leq \gamma \max_{s\in S, a\in A} \|\widehat{T}(s,a,\cdot)\|_1 \|V_{m-1} - V_{m-2}\|_\infty \\
&= \gamma \|V_{m-1} - V_{m-2}\|_\infty \leq \gamma \|Q_{m-1} - Q_{m-2}\|_\infty,
\end{aligned}$$

so

$$\|Q_m - Q_0\|_\infty \leq \sum_{k=0}^{m-1} \|Q_{k+1} - Q_k\|_\infty \leq \|Q_1 - Q_0\|_\infty \sum_{k=1}^{m-1} \gamma^{k-1}.$$

Taking the limit of $m \to \infty$, $Q_m \to Q^{\pi}_{\widehat{M},\gamma}$, and we have

$$\left\|Q^{\pi}_{\widehat{M},\gamma} - Q_0\right\|_\infty \leq \frac{1}{1-\gamma} \|Q_1 - Q_0\|_\infty.$$

This completes the proof, noticing that $Q_0 = Q^{\pi}_{M,\gamma}$, $V_0 = V^{\pi}_{M,\gamma}$, and $Q_1(s,a) = \widehat{R}(s,a) + \gamma\langle \widehat{T}(s,a,\cdot), V^{\pi}_{M,\gamma}\rangle$. □

Proof of Lemma 2 From Equation 18 in Lemma 3 and Lemma 4, we have

$$\begin{aligned}
&\left\|V^{\pi^*_{M,\gamma}}_{M,\gamma} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma}\right\|_\infty \leq 2 \max_{\pi\in\Pi_{R,\gamma}} \left\|V^{\pi}_{M,\gamma} - V^{\pi}_{\widehat{M},\gamma}\right\|_\infty \\
&\leq 2 \max_{\pi\in\Pi_{R,\gamma}} \left\|Q^{\pi}_{M,\gamma} - Q^{\pi}_{\widehat{M},\gamma}\right\|_\infty \\
&= 2 \max_{\substack{s\in S, a\in A \\ \pi\in\Pi_{R,\gamma}}} \left|Q^{\pi}_{M,\gamma}(s,a) - Q^{\pi}_{\widehat{M},\gamma}(s,a)\right| \\
&\leq \frac{2}{1-\gamma} \max_{\substack{s\in S, a\in A \\ \pi\in\Pi_{R,\gamma}}} \left|R(s,a) + \gamma\langle \widehat{T}(s,a,\cdot), V^{\pi}_{M,\gamma}\rangle - Q^{\pi}_{M,\gamma}(s,a)\right|.
\end{aligned}$$

For any particular s, a, π tuple, according to Hoeffding's inequality, $\forall t > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\left|R(s,a) + \gamma\langle \widehat{T}(s,a,\cdot), V^{\pi}_{M,\gamma}\rangle - Q^{\pi}_{M,\gamma}(s,a)\right| > t\right\} \\
&\leq 2\exp\left\{-\frac{2nt^2}{R^2_{\max}/(1-\gamma)^2}\right\}, \qquad (19)
\end{aligned}$$

as $(R(s,a) + \gamma\langle \widehat{T}(s,a,\cdot), V^{\pi}_{M,\gamma}\rangle)$ is the average of i.i.d. samples bounded in $[0, R_{\max}/(1-\gamma)]$, with mean $Q^{\pi}_{M,\gamma}(s,a)$. To obtain a uniform bound over all (s,a,π) tuples, we set the right-hand side of Equation 19 to $\delta/|S||A||\Pi_{R,\gamma}|$, and solve for t, and the theorem follows.

Proof of Theorem 2. $\forall s \in S$,

$$\begin{aligned}
V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma_{\text{eval}}}(s) - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{\text{eval}}}(s) = &\left(V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma_{\text{eval}}}(s) - V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma}(s)\right) \\
&+ \left(V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma}(s) - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{\text{eval}}}(s)\right).
\end{aligned}$$

By Lemma 1, the first term can be bounded by

$$V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma_{\text{eval}}}(s) - V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma}(s) \leq \frac{\gamma_{\text{eval}} - \gamma}{(1-\gamma_{\text{eval}})(1-\gamma)} R_{\max}$$

and by Lemma 2, the second term can be bounded as follows w.p. at least $1-\delta$:

$$\begin{aligned}
&V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma}(s) - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{\text{eval}}}(s) \leq V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma}(s) - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma}(s) \\
&\leq V^{\pi^*_{M,\gamma}}_{M,\gamma}(s) - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma}(s) \qquad (\pi^*_{M,\gamma} \text{ is optimal for } (M,\gamma)) \\
&\leq \frac{2R_{\max}}{(1-\gamma)^2}\sqrt{\frac{1}{2n}\log\frac{2|S||A||\Pi_{R,\gamma}|}{\delta}}.
\end{aligned}$$

B. PROOF OF THEOREM 3

We prove Theorem 3 by the following lemma that parallels Lemma 2.

Lemma 5. *Given the true MDP M, let $\widehat{M}$ be an MDP comprising reward function $\widehat{R}$ and transition function $\widehat{T}$ both estimated from n samples for each state-action pair, then*

$$\begin{aligned}
&\left\|V^{\pi^*_{M,\gamma}}_{M,\gamma} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma}\right\|_\infty \qquad (20) \\
&\leq \frac{2}{1-\gamma}\left(2\max_{\substack{s\in S \\ a\in A}} \widehat{\mathfrak{R}}_{D_{s,a}}(\mathcal{F}_{M,\gamma}) + \frac{3R_{\max}}{1-\gamma}\sqrt{\frac{1}{2n}\log\frac{4|S||A|}{\delta}}\right),
\end{aligned}$$

with probability at least $1-\delta$.

PROOF. From Equation 17 in Lemma 3 and Lemma 4, we have

$$\begin{aligned}
&\left\|V^{\pi^*_{M,\gamma}}_{M,\gamma} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma}\right\|_\infty \\
&\leq \frac{2}{1-\gamma} \max_{\substack{s\in S, a\in A \\ \pi: S\to A}} \left|\widehat{R}(s,a) + \gamma\langle \widehat{T}(s,a,\cdot), V^{\pi}_{M,\gamma}\rangle - Q^{\pi}_{M,\gamma}(s,a)\right| \\
&= \frac{2}{1-\gamma} \max_{s\in S, a\in A} \max_{\pi: S\to A} \left|\widehat{R}(s,a) + \gamma\langle \widehat{T}(s,a,\cdot), V^{\pi}_{M,\gamma}\rangle - Q^{\pi}_{M,\gamma}(s,a)\right|.
\end{aligned}$$

Recall that in the statement of Theorem 3, we defined $f^{\pi}_{M,\gamma}$ to be the mapping $(r,s') \mapsto r + \gamma V^{\pi}_{M,\gamma}(s')$. So

$$\begin{aligned}
&\max_{\pi: S\to A} \left|\widehat{R}(s,a) + \gamma\langle \widehat{T}(s,a,\cdot), V^{\pi}_{M,\gamma}\rangle - Q^{\pi}_{M,\gamma}(s,a)\right| \\
&= \max_{\pi: S\to A} \left|\frac{1}{n}\sum_{(r,s')\in D_{s,a}} f^{\pi}_{M,\gamma}(r,s') - \mathop{\mathbb{E}}_{(r,s')\sim\mathbb{P}_{s,a}}\left\{f^{\pi}_{M,\gamma}(r,s')\right\}\right|,
\end{aligned}$$

where $(r,s') \in D_{s,a}$ means that (r,s') is a sample reward & next-state pair from (s,a) in dataset D, and $\mathbb{P}_{s,a}$ is the underlying true distribution. By noticing that $f^{\pi}_{M,\gamma}$ has function value bounded in $[0, R_{\max}/(1-\gamma)]$, we have the following bound from the standard Rademacher complexity literature (e.g., [3]; also see [21]): for each $s \in S, a \in A$, w.p. $\geq 1 - \delta/(|S||A|)$,

$$\begin{aligned}
&\max_{\pi: S\to A} \left|\frac{1}{n}\sum_{(r,s')\in D_{s,a}} f^{\pi}_{M,\gamma}(r,s') - \mathop{\mathbb{E}}_{(r,s')\sim\mathbb{P}_{s,a}}\left\{f^{\pi}_{M,\gamma}(r,s')\right\}\right| \\
&\leq \frac{2}{1-\gamma}\left(2\widehat{\mathfrak{R}}_{D_{s,a}}(\mathcal{F}_{M,\gamma}) + \frac{3R_{\max}}{1-\gamma}\sqrt{\frac{1}{2n}\log\frac{4|S||A|}{\delta}}\right).
\end{aligned}$$

And the theorem follows directly from union bound and taking the maximal empirical Rademacher complexity among all state-action pairs. □

REFERENCES

[1] Michael Kearns, Yishay Mansour, and Andrew Y Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning*, 49(2-3):193–208, 2002.

[2] Levente Kocsis and Csaba Szepesvári. Bandit based Monte-Carlo planning. In *Machine Learning: ECML 2006*, pages 282–293. 2006.

[3] Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[4] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012.

[5] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838, 1992.

[6] M.J. Kearns and U.V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

[7] Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error Propagation for Approximate Policy and Value Iteration. In *Advances in Neural Information Processing Systems*, pages 568–576, 2010.

[8] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

[9] Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite MDPs: PAC analysis. *The Journal of Machine Learning Research*, 10:2413–2444, 2009.

[10] Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. In *Learning Theory*, pages 157–171. Springer, 2007.

[11] Xiaojin Zhu, Bryan R Gibson, and Timothy T Rogers. Human Rademacher complexity. In *Advances in Neural Information Processing Systems*, pages 2322–2330, 2009.

[12] Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–457. Springer, 2000.

[13] Nan Jiang, Satinder Singh, and Richard Lewis. Improving UCT planning via approximate homomorphisms. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1289–1296, 2014.

[14] David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. In *Advances in Neural Information Processing Systems*, pages 2164–2172, 2010.

[15] Trey Smith and Reid Simmons. Heuristic search value iteration for POMDPs. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 520–527, 2004.

[16] Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

[17] Cosmin Paduraru. *Off-policy Evaluation in Markov Decision Processes*. PhD thesis, McGill University, 2013.

[18] Balaraman Ravindran and Andrew Barto. Approximate homomorphisms: A framework for nonexact minimization in Markov decision processes. In *Proceedings of the 5th International Conference on Knowledge-Based Computer Systems*, 2004.

[19] Ambuj Tewari and Peter L Bartlett. Sample complexity of policy search with known dynamics. In *Advances in Neural Information Processing Systems*, volume 19, pages 97–104, 2006.

[20] Marek Petrik and Bruno Scherrer. Biasing approximate dynamic programming with a lower discount factor. In *Advances in Neural Information Processing Systems*, pages 1265–1272, 2009.

[21] Maria-Florina Balcan. *CS 8803 - Machine Learning Theory: Lecture Notes*. Georgia Institute of Technology, 2011. `http://www.cc.gatech.edu/~ninamf/ML11/lect1115.pdf`.