On Structural Properties of MDPs that Bound Loss due to Shallow Planning

Nan Jiang[1] and **Satinder Singh**[1] and **Ambuj Tewari**[2]
[1]Computer Science and Engineering, University of Michigan
[2]Department of Statistics, University of Michigan
{nanjiang, baveja, tewaria}@umich.edu

Abstract

Planning in MDPs often uses a smaller planning horizon than specified in the problem to save computational expense at the risk of a loss due to suboptimal plans. Jiang *et al.* [2015b] recently showed that smaller than specified planning horizons can in fact be beneficial in cases where the MDP model is learned from data and therefore not accurate. In this paper, we consider planning with *accurate* models and investigate structural properties of MDPs that *bound* the loss incurred by using smaller than specified planning horizons. We identify a number of structural parameters some of which depend on the reward function alone, some on the transition dynamics alone, and some that depend on the interaction between rewards and transition dynamics. We provide planning loss bounds in terms of these structural parameters and, in some cases, also show tightness of the upper bounds. Empirical results with randomly generated MDPs are used to validate qualitative properties of our theoretical bounds for shallow planning.

1 Introduction

Planning in Markov Decision Processes (MDPs) involves a lookahead at the consequences of potential action choices using a computational-model of the transition dynamics and the reward function components of the MDP. The horizon specified as part of the planning problem determines how deep (far into the future) the lookahead has to be. The longer the planning-horizon the greater the computational effort needed to compute an optimal policy. To save on this computational effort, planners often use a smaller than specified planning horizon; hereafter we refer to this as *shallow planning*. Of course, the computation saved by shallow planning comes at the cost of obtaining a policy that is suboptimal relative to the optimal policy. Recent work by Jiang *et al.* [2015b] shows that when planning with *inaccurate* models (perhaps learned from small amounts of data) it can actually be beneficial to use shallow planning because it avoids overfitting to the noise in the inaccurate model. In this paper, we focus exclusively on the setting of planning with *accurate* models with the goal of understanding what properties of the MDP help determine the loss due to shallow planning.

A widely understood but coarse upper bound (see Equation 2) on the loss due to shallow planning is outlined below; it is based only on the largest reward in the MDP and does not exploit any other finer-grained properties of the MDP. In this paper, we identify a set of structural properties of an MDP some of which depend on the reward function alone, some on the transition dynamics alone, and some that depend on the interaction between rewards and transition dynamics. We provide planning loss bounds in terms of these structural parameters and, in some cases, also show tightness of the upper bounds. Empirical results with randomly generated MDPs are used to validate qualitative properties of our theoretical bounds for shallow planning.

2 Planning Setting & Notation

An MDP is specified as a tuple $M = \langle S, A, P, R, \gamma_{\text{eval}} \rangle$ where S is the state space; A is the action space; $P : S \times A \times S \to [0, 1]$ is the transition function, and $R : S \times A \to [0, R_{\max}]$ is the reward function. The evaluation discount factor, $\gamma_{\text{eval}} \in [0, 1)$, determines the effective planning horizon of the problem (more on this below). The planning task is to compute an optimal policy, a mapping from states to actions, that maximizes value, that is the expected sum of future rewards discounted by γ_{eval} at every time step. Given a policy $\pi : S \to A$, we use $V^{\pi}_{M,\gamma_{\text{eval}}}(s)$ to denote its value as a function of the starting state s (with a slight abuse of notation we will also treat $V^{\pi}_{M,\gamma_{\text{eval}}}$ as a vector in $\mathbb{R}^{|S|}$). Given an MDP, there always exists a policy that simultaneously maximizes the value of all states, and we denote such an optimal policy as $\pi^*_{M,\gamma_{\text{eval}}}$.

Throughout this paper we will use the phrases "discount factor" and "planning horizon" interchangeably, since the infinite sum of rewards discounted by γ_{eval} are approximated by a finite horizon of order $O(1/(1-\gamma_{\text{eval}}))$ in online planning algorithms such as Monte-carlo Tree Search methods [Kearns *et al.*, 2002]. In practice, to save computational costs a shallow planner would use a discount factor $\gamma < \gamma_{\text{eval}}$ to guide the planning algorithm in its computation/search of a good policy. We emphasize that the ultimate goodness of the policy π found by the shallow planner will still be evaluated in M using γ_{eval}. To facilitate analysis we ignore details of specific shallow planning algorithms and instead assume perfect

planning under γ, i.e., we assume the shallow planning algorithm outputs $\pi^*_{M,\gamma}$, the policy that is optimal in M for a discount factor γ. We define the **loss** due to shallow planning as the worst (over all states) absolute difference in value of the optimal policy $\pi^*_{M,\gamma_{\text{eval}}}$ and $\pi^*_{M,\gamma}$, i.e.,

$$\left\| V^{\pi^*_{M,\gamma_{\text{eval}}}}_{M,\gamma_{\text{eval}}} - V^{\pi^*_{M,\gamma}}_{M,\gamma_{\text{eval}}} \right\|_\infty . \quad (1)$$

Finally, because we only consider planning with accurate models, hereafter we drop explicit dependence on M in all notation (for value functions and policies) unless otherwise specified, for they are all automatically with reference to the true MDP M.

3 Parameters that Bound Loss

Before we turn to our finer-grained parameters that bound loss due to shallow planning, we note here that there is a straightforward bound on loss that comes simply from the largest reward (this has been explicitly given by Petrik & Scherrer [2009]; we derive it in Section 3.2 for completeness):

$$\left\| V^{\pi^*_{\gamma_{\text{eval}}}}_{\gamma_{\text{eval}}} - V^{\pi^*_{\gamma}}_{\gamma_{\text{eval}}} \right\|_\infty \leq \frac{\gamma_{\text{eval}} - \gamma}{(1-\gamma_{\text{eval}})(1-\gamma)} R_{\max}. \quad (2)$$

This bound ignores the role of transitions in determining value functions, and indeed any other aspect of reward functions but its largest value, and finally any interaction between rewards and transitions.

3.1 Value Function Variation Parameter & Loss

We begin with a summary "intermediate" parameter and a bound on the loss derived from it; later we show how to relate this summary parameter to several high-level properties of an MDP that pertain to rewards, transitions, and their interaction.

Definition 1. $\kappa_\gamma = \max_{s,s' \in S} \left| V^{\pi^*_\gamma}_\gamma(s) - V^{\pi^*_\gamma}_\gamma(s') \right|$,

For (environment M and) discount factor γ the value-function-variation parameter, κ_γ, measures the maximal variation in optimal value between any two states, or equivalently how much difference the choice of start state can make on the value an agent can achieve. Note that the quantities in κ_γ depend only on γ and not at all on γ_{eval}. Next we use this parameter to bound the loss defined in Equation 1 as follows.

Theorem 1 (Upper-Bound on Loss from κ_γ)**.**

$$\left\| V^{\pi^*_{\gamma_{\text{eval}}}}_{\gamma_{\text{eval}}} - V^{\pi^*_{\gamma}}_{\gamma_{\text{eval}}} \right\|_\infty \leq \frac{\gamma_{\text{eval}} - \gamma}{1-\gamma_{\text{eval}}} \kappa_\gamma.$$

Proof. For any policy π, we can write $V^\pi_{\gamma_{\text{eval}}}$ as a linear combination of V^π_γ by decomposing and re-arranging the reward obtained at each step (see Appendix A.1 for details of this step)[1]: $\forall s \in S$, let e_s be the unit vector with the element indexed by s equal to 1, we have

$$\begin{aligned} V^\pi_{\gamma_{\text{eval}}}(s) &= e_s^\top \textstyle\sum_{t=1}^\infty {\gamma_{\text{eval}}}^{t-1} [P^\pi]^{t-1} R^\pi \\ &= e_s^\top V^\pi_\gamma + \textstyle\sum_{k=1}^\infty (\gamma_{\text{eval}} - \gamma) {\gamma_{\text{eval}}}^{k-1} e_s^\top [P^\pi]^k V^\pi_\gamma, \end{aligned}$$

where $[P^\pi]$ is a $|S| \times |S|$ matrix with the element indexed by (s, s') being $P(s'|s, a)$, and R^π is a $|S| \times 1$ vector with the s-th element being $R(s, \pi(s))$. Now plug in π^*_γ and $\pi^*_{\gamma_{\text{eval}}}$ and consider the value difference: $\forall s \in S$,

$$\begin{aligned} &V^{\pi^*_{\gamma_{\text{eval}}}}_{\gamma_{\text{eval}}}(s) - V^{\pi^*_\gamma}_{\gamma_{\text{eval}}}(s) = \left(e_s^\top V^{\pi^*_{\gamma_{\text{eval}}}}_\gamma - e_s^\top V^{\pi^*_\gamma}_\gamma\right) + \\ &(\gamma_{\text{eval}} - \gamma) \sum_{k=1}^\infty \gamma_{\text{eval}}^{k-1} \left(e_s^\top [P^{\pi^*_{\gamma_{\text{eval}}}}]^k V^{\pi^*_{\gamma_{\text{eval}}}}_\gamma - e_s^\top [P^{\pi^*_\gamma}]^k V^{\pi^*_\gamma}_\gamma\right) \\ &\leq e_s^\top \left(V^{\pi^*_{\gamma_{\text{eval}}}}_\gamma - V^{\pi^*_\gamma}_\gamma\right) + \quad (3) \\ &(\gamma_{\text{eval}} - \gamma) \sum_{k=1}^\infty \gamma_{\text{eval}}^{k-1} \left(e_s^\top [P^{\pi^*_{\gamma_{\text{eval}}}}]^k V^{\pi^*_\gamma}_\gamma - e_s^\top [P^{\pi^*_\gamma}]^k V^{\pi^*_\gamma}_\gamma\right) \\ &\leq 0 + (\gamma_{\text{eval}} - \gamma) \textstyle\sum_{k=1}^\infty \gamma_{\text{eval}}^{k-1} \kappa_\gamma = \frac{\gamma_{\text{eval}} - \gamma}{1-\gamma_{\text{eval}}} \kappa_\gamma. \end{aligned}$$

The first inequality above holds by optimality of π^*_γ for discount factor γ and the second inequality holds because for any stochastic vectors p, q, we have $|p^\top V^{\pi^*_\gamma}_\gamma - q^\top V^{\pi^*_\gamma}_\gamma| \leq \kappa_\gamma$. □

3.2 Bounding Value Function Variation using other parameters

Here we develop several structural parameters that bound κ_γ.

Rewards-Only Parameters

The first rewards-only parameter is simply the largest reward. Proposition 1 below shows how it can be used to bound κ_γ (the proof is straightforward and hence omitted) and when this is applied to Theorem 1 we get the known bound on loss due to shallow planning in Equation 2.

Proposition 1. $\kappa_\gamma \leq R_{\max}/(1-\gamma)$.

The next rewards-only parameter stems from the observation that if we were able to obtain $R_{\max}$ immediate reward in every state, there would be no need to plan ahead. In general, we show that the loss of shallow planning can be bounded by the extent to which this criterion is violated.

Definition 2 (Reward Variation)**.**
$\Delta_R = \max_{s,s' \in S} |\max_a R(s,a) - \max_{a'} R(s',a')|$.

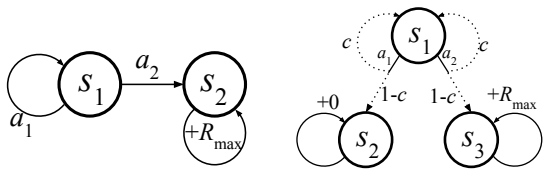
Proposition 2. $\kappa_\gamma \leq \Delta_R/(1-\gamma)$.

Proof. The myopic policy $s \mapsto \text{argmax}_a R(s,a)$ yields at least $\min_s \max_a R(s,a)/(1-\gamma)$ value from any starting state, which is a lower bound on $V^{\pi^*_\gamma}_\gamma$. On the other hand, any policy and starting state pair cannot have value more than $\max_{s,a} R(s,a)/(1-\gamma)$, and the proposition follows by combining the two bounds. □

Worst case tightness We show that the planning loss bound based on Δ_R by combining Theorem 1 and Proposition 2 is tight in the worst case (and so is Theorem 1 itself, as a direct corollary).

Claim 1. *For any $\Delta_R \in [0, R_{\max}]$, $\gamma \in [0, 1 - \Delta_R/R_{\max}]$, $\gamma_{\text{eval}} \in [\gamma, 1)$, there exists an MDP M with reward variation Δ_R, and the loss incurred by using γ is equal to the bound given by Theorem 1 and Proposition 2.*

[1] All appendices mentioned in this paper are included in an extended version available at `https://sites.google.com/a/umich.edu/nanjiang/ijcai2016-horizon.pdf`.



(a) Worst-case MDP for proving Claim 1

(b) Worst-case MDP for proving Claim 2

Figure 1: MDPs constructed to prove Claim 1 and 2. In both cases, a_1 is optimal under γ and a_2 is optimal under γ_{eval}. **(a)** $R(s_1, a_1) = R_{\max} - \Delta_R$ and $R(s_1, a_2) = R_{\max} - \Delta_R/(1-\gamma)$. **(b)** Dotted arrows represent stochastic transitions, and $c = 1 - \delta_P/2$. $R(s_1, a_1) = 0$ and $R(s_1, a_2) = \gamma\delta_P R_{\max}/2(1-\gamma)$.

Corollary 1. *For any $\kappa_\gamma \in [0, R_{\max}], \gamma \in [0, 1), \gamma_{\text{eval}} \in [\gamma, 1)$, there exists an MDP with value function variation κ_γ and the loss incurred by using γ is equal to the bound given by Theorem 1.*

Claim 1 is proved by constructing a two-state MDP illustrated in Figure 1a, and proof details are deferred to Appendix A.2.

Transitions-Only Parameters

By using structure in the transition probabilities we can get tighter bounds on κ_γ. The next parameter, ϵ-mixing time, is motivated by the fact that if the MDP mixes fast under a policy π, then the value function of π has small variation over the state space. When $\pi = \pi^*_\gamma$, the parameter yields a bound on κ_γ (explicitly stated in Corollary 2). For ease of technical presentation we will assume that the Markov Chain induced by all policies we consider is ergodic (i.e., under any policy, it is possible to reach any state from any other state).

Definition 3 (ϵ-mixing time)**.** *Define the ϵ-mixing time for policy π as*

$$T_\pi(\epsilon) = \inf\left\{T : \forall s \in S, t \geq T, \left\|e_s^\top [P^\pi]^t - (\rho^\pi)^\top\right\|_1 \leq \epsilon\right\}$$

where ρ^π is the limiting distribution independent of the starting state.

Proposition 3. *For policy π,*

$$\max_{s,s' \in S} \left|V_\gamma^\pi(s) - V_\gamma^\pi(s')\right| \leq \tfrac{R_{\max}}{1-\gamma}\left(1 - \gamma^{T_\pi(\epsilon)} + \epsilon\gamma^{T_\pi(\epsilon)}\right).$$

Corollary 2. $\kappa_\gamma \leq R_{\max}\left(1 - \gamma^{T_{\pi^*_\gamma}(\epsilon)} + \epsilon\gamma^{T_{\pi^*_\gamma}(\epsilon)}\right)/(1-\gamma)$.

Proposition 3 can be proved by simply reducing to Proposition 5 after observing that $T_\pi(\epsilon)$ is a $(\epsilon R_{\max}/2)$-return mixing time for π as defined in Definition 5. The actual proof is deferred to Appendix A.3.

The next transition-only parameter is the stochastic diameter T_M, the longest expected time to travel from one state to another. If this parameter is small, $V_\gamma^{\pi^*_\gamma}$ must have a small variation, otherwise we could improve the value of low-valued states by a non-stationary policy that travels to a high-valued state first and executes the optimal policy afterwards.

Definition 4 (Stochastic diameter)**.**

$$T_M = \max_{s,s' \in S} \min_{\pi: S \to A} \mathbb{E}\left\{\inf\left\{t \in \mathbb{N} : s_t = s'\right\} \middle| s_0 = s, \pi\right\}.$$

Proposition 4. $\kappa_\gamma \leq \frac{1-\gamma^{T_M}}{1-\gamma} R_{\max}$.

Proof. It suffices to show that for any $s, s' \in S$, $V_\gamma^{\pi^*_\gamma}(s') - V_\gamma^{\pi^*_\gamma}(s) \leq \frac{1-\gamma^{T_M}}{1-\gamma} R_{\max}$. Since π^*_γ is optimal under γ, we can lower bound $V_\gamma^{\pi^*_\gamma}(s)$ by the value obtained by starting at s and following any policy. In particular, consider a non-stationary policy that first travels to s' by executing the policy that achieves the minimum in the definition of T_M, and then switch to π^*_γ. Suppose it takes t steps to get to s' (t is a random variable), then the non-stationary policy gives at least $\gamma^t V_\gamma^{\pi^*_\gamma}(s')$ value, and

$$\begin{aligned} V_\gamma^{\pi^*_\gamma}(s) &\geq \mathbb{E}\left\{\gamma^t V_\gamma^{\pi^*_\gamma}(s')\right\} = \mathbb{E}\left\{\gamma^t\right\} V_\gamma^{\pi^*_\gamma}(s') \\ &\geq \gamma^{\mathbb{E}[t]} V_\gamma^{\pi^*_\gamma}(s') \qquad (f(x) = \gamma^x \text{ is convex}) \\ &\geq \gamma^{T_M} V_\gamma^{\pi^*_\gamma}(s') \geq V_\gamma^{\pi^*_\gamma}(s') - \tfrac{1-\gamma^{T_M}}{1-\gamma} R_{\max}. \quad \square \end{aligned}$$

Transitions-and-Rewards Parameters

Thus far, we have provided parameters of rewards alone and transitions alone. Here we consider a parameter that captures the interaction between rewards and transitions, the ϵ-return mixing time, which measures mixing via the closeness of the expected reward obtained after a particular time step and that obtained in the long run.[2]

Definition 5 (ϵ-return mixing time)**.**

$$T_\pi^v(\epsilon) = \inf\left\{T : \forall t \geq T, \ \left\|[P^\pi]^t R^\pi - \eta^\pi\right\|_\infty \leq \epsilon\right\},$$

where scalar η^π is the average reward per step of policy π.

Proposition 5.

$$\max_{s,s' \in S} \left|V_\gamma^\pi(s) - V_\gamma^\pi(s')\right| \leq \frac{R_{\max}(1 - \gamma^{T_\pi^v(\epsilon)}) + 2\epsilon\gamma^{T_\pi^v(\epsilon)}}{1-\gamma}.$$

Corollary 3. $\kappa_\gamma \leq \left(R_{\max}(1 - \gamma^{T_{\pi^*_\gamma}^v(\epsilon)}) + 2\epsilon\gamma^{T_{\pi^*_\gamma}^v(\epsilon)}\right)/(1-\gamma)$.

Proof Sketch of Proposition 5. (full proof in Appendix A.4) Recall that $V_\gamma^\pi = \sum_{t=1}^\infty \gamma^{t-1}[P^\pi]^{t-1} R^\pi$. We keep the first $T_\pi^v(\epsilon)$ terms in the summation, and approximate the rest of the terms by $\gamma^{t-1}\eta^\pi$. Let $[V_\gamma^\pi]'$ denote such an approximation. By the definition of $T_\pi^v(\epsilon)$, $|V_\gamma^\pi - [V_\gamma^\pi]'| \leq \epsilon\gamma^{T_\pi^v(\epsilon)}/(1-\gamma)$. On the other hand, for any $s, s' \in S$, $[V_\gamma^\pi]'(s) - [V_\gamma^\pi]'(s')$ only differ in the first $T_\pi^v(\epsilon)$ terms in the expansion, hence the difference can be bounded by $\frac{1-\gamma^{T_\pi^v(\epsilon)}}{1-\gamma} R_{\max}$. The proposition follows by combining the two sources of difference. $\square$

4 Action Variation

In the preceding section, we looked at structural parameters of the MDP that bound the loss due to shallow planning, all via an intermediate quantity κ_γ that characterizes the value function variation. Yet, there are MDPs with large κ_γ that still have a small loss. While covering all such cases is outside the scope of this paper, we cover one such class of MDPs, A natural class of such MDPs are those where different actions at

[2] This definition is slightly adapted from Kearns & Singh [2002], which considered the *average* reward obtained in first T time steps.

the same state have almost identical distributions over next-states; no deep planning is needed in these MDPs as they are essentially contextual bandits (except that the contexts are not i.i.d.). We capture this idea by the notion of Action Variation, and provide an associated loss bound which subsumes Theorem 1.

Definition 6 (Action Variation)**.**

$$\delta_P = \max_{s\in S}\max_{a,a'\in A} \|P(\cdot|s,a) - P(\cdot|s,a')\|_1 .$$

Theorem 2.

$$\left\| V_{\gamma_{\text{eval}}}^{\pi^*_{\gamma_{\text{eval}}}} - V_{\gamma_{\text{eval}}}^{\pi^*_{\gamma}} \right\|_\infty \leq \frac{\delta_P/2 \cdot \kappa_\gamma(\gamma_{\text{eval}} - \gamma)}{(1-\gamma_{\text{eval}})(1-\gamma_{\text{eval}}(1-\delta_P/2))}.$$

Theorem 2 is a planning loss bound that depends on both δ_P and κ_γ; the bound monotonically increases with δ_P and reduces to Theorem 1 when δ_P takes the maximal value 2.

To prove the theorem, we first define the *commonality* between two probability distributions, and state a key lemma w.r.t. this quantity. Note that commonality appears in the mixing time literature for Markov chains in linking the notions of total variation and coupling [Levin *et al.*, 2009, Section 4.2].

Definition 7. *Given two vectors p, q of the same dimension, define $\text{comm}(p,q)$ as the commonality vector of p and q, whose s-th element is $\text{comm}(s;\ p,q) = \min\{p(s), q(s)\}$.*

Fact 1. *When p and q are stochastic vectors,*

$$\|\text{comm}(p,q)\|_1 = 1 - \|p-q\|_1/2.$$

Lemma 1. *Suppose p and q are stochastic vectors over S, $\forall \pi_1, \pi_2 : S \to A$,*

$$\|\text{comm}(p^\top P^{\pi_1}, q^\top P^{\pi_2})\|_1 \geq (1-\delta_P/2)\|\text{comm}(p,q)\|_1.$$

Corollary 4. *For any $s \in S, \pi_1, \pi_2 : S \to A, k \in \mathbb{N}$,*

$$\|e_s^\top [P^{\pi_1}]^k - e_s^\top [P^{\pi_2}]^k\|_1 \leq 2 - 2(1-\delta_P/2)^k.$$

Proof. Use Fact 1 to turn ℓ_1 error into commonality, apply Lemma 1 k times, and notice that $\|\text{comm}(e_s, e_s)\|_1 = 1$. □

Theorem 2 follows straightforwardly from applying Corollary 4 to Equation 3 in the proof of Theorem 1. We only include the proof to the key lemma below, and the proof of Theorem 2 is deferred to Appendix A.5.

Proof of Lemma 1. Let $P^\pi(s|\cdot)$ be a column vector of transition probabilities from each state to s under policy π, then

$$\begin{aligned}
&\text{comm}(s;\ p^\top P^{\pi_1}, q^\top P^{\pi_2}) = \min\{p^\top P^{\pi_1}(s|\cdot), q^\top P^{\pi_2}(s|\cdot)\} \\
&\geq \min\{\text{comm}(q,p)^\top P^{\pi_1}(s|\cdot), \text{comm}(q,p)^\top P^{\pi_2}(s|\cdot)\} \\
&= \text{comm}(s;\ \text{comm}(p,q)^\top P^{\pi_1}, \text{comm}(p,q)^\top P^{\pi_2}).
\end{aligned}$$

Define z as $\text{comm}(p,q)$ normalized so that $\|z\|_1 = 1$, then

$$\begin{aligned}
&\|\text{comm}(p^\top P^{\pi_1}, q^\top P^{\pi_2})\|_1 \\
&\geq \|\text{comm}(\text{comm}(p,q)^\top P^{\pi_1}, \text{comm}(p,q)^\top P^{\pi_2})\|_1 \\
&= \|\text{comm}(p,q)\|_1 \|\text{comm}(z^\top P^{\pi_1}, z^\top P^{\pi_2})\|_1 \\
&= \|\text{comm}(p,q)\|_1 (1 - \|z^\top (P^{\pi_1} - P^{\pi_2})\|_1/2) \quad \text{(Fact 1)} \\
&\geq \|\text{comm}(p,q)\|_1 (1-\delta_P/2).
\end{aligned}$$

The last step uses the fact that $\|\cdot\|_1$ is a convex function, and each row of $P^{\pi_1} - P^{\pi_2}$ has ℓ_1-norm bounded by δ_P. □

Worst case tightness We show that Theorem 2 is tight in the worst case, as we did for Proposition 2.

Claim 2. *For any $\delta_P \in [0,2]$, $\gamma \in [0, 1/(1+\delta_P/2)]$, $\gamma_{\text{eval}} \in [\gamma, 1)$, there exists an MDP M with Action Variation equal to δ_P and a planning loss of using γ equal to the bound given in Theorem 2.*

The claim is proved by constructing a 3-state MDP illustrated in Figure 1b, and the proof is deferred to Appendix A.6.

5 Empirical Illustrations

Our theoretical results translate various structural properties of an MDP onto a smooth upper-bound on the loss due to shallow planning. But what does the actual loss look like in any particular MDP? We know that the loss curve as a function of γ should be piecewise constant. This is because as we lower γ from γ_{eval} towards zero, there will be discrete points at which the optimal policy with respect to γ changes, partitioning the discount-factor interval and yielding a piecewise constant loss curve. This behavior of the loss curve is consistent with Blackwell optimality [Hordijk and Yushkevich, 2002], which asserts that at the extreme end near $\gamma = 1$ the loss curve is constant with value 0. This is seen in the 4 panels of Figure 2 where we plot the loss as a function of γ (see caption for additional details; we describe how the specific MDPs were generated below). What is perhaps interesting is that the loss curves are not always non-decreasing with increasing γ. The loss-curve in the bottom-right panel is clearly non-monotonic. The graph on the right of Figure 2 shows a simple MDP where it is easy to see how the loss can be non-monotonic as a function of γ. Thus, loss-curves as a function of γ in any specific MDP can be complex and hard to predict using only high-level structural properties.

A useful way to illustrate the empirical validity of our monotonic theoretical results is to consider "average" loss curves by sampling MDPs from some distribution. Intuitively averaging multiple piecewise constant loss curves from perturbed MDPs should yield smooth loss curves (this is a form of smoothed analysis [Spielman and Teng, 2009]). Specifically, the results presented below will be of the following form. We will sample MDPs from multiple different generative-distributions defined in Section 5.1. Using procedures defined in Section 5.2, for each random MDP we will compute the empirical value of the loss and the structural properties defined in Section 3.2 and 4. Then to show that the structural properties matter we group their values into quantiles and plot an average loss curve for each quantile by averaging the loss curves over the MDPs that fall into that quantile. As we discuss below we get the qualitative phenomenon expected from our theoretical results.

5.1 Domains Specification

We consider random MDPs with N states and 2 actions, generated according to the following schemes.

1. Random topologies: Each state-action pair is randomly assigned d possible next-states, where d is chosen according to one of the following:

   `fixed(N,d)`: d is a fixed number.

`binom(N,p)`: d is binomially distributed as $B(N, p)$.

2. Ring topology `ring(N,p)`: the N states form a ring. Upon taking action 1 at a state, the agent either stays at the same place or moves to the next state in clockwise order; the same for action 2 except that the agent moves in a counter-clockwise order. In addition, for each (s, a, s') where s is not next to s', with probability p we add s' as a next-state for (s, a).

Once the connectivity-structure of an MDP is determined as above, we fill the non-zero entries in the transition probabilities and the rewards with numbers drawn independently from $U[0, 1]$ and normalize the transition probabilities.

5.2 Computing Structural Parameters

We compute the quantities that our theoretical results refer to for every random MDP that we generate. Below is the list of quantities and how we compute them in practice:

1. Relative loss:

$$\max_{s\in S}\left(V^{\pi^*_{\gamma_{\text{eval}}}}_{\gamma_{\text{eval}}}(s) - V^{\pi^*_\gamma}_{\gamma_{\text{eval}}}(s)\right)/V^{\pi^*_{\gamma_{\text{eval}}}}_{\gamma_{\text{eval}}}(s).$$

   This is the empirical version of Equation 1, with a normalized magnitude between $[0, 1]$.

2. Reward variation: we use $\Delta_R/\max_{s,a} R(s, a)$ as the empirical version of Δ_R.

3. ϵ-mixing time: we compute $T_\pi(\epsilon)$ by its definition in Proposition 3 with $\epsilon = 0.01, \pi = \pi^*_{\gamma_{\text{eval}}}$. An implementation detail is that, we only search 50 steps for $T_\pi(\epsilon)$ instead of checking an infinite number of steps, which is sufficient for the MDP distributions we consider in this paper. Formally, the empirical version of $T_\pi(\epsilon)$ is $\inf\left\{T \le 50 : \forall s \in S, T \le t \le 50,\ \left\|e_s^\top [P^\pi]^t - (\rho^\pi)^\top\right\|_1 \le \epsilon\right\}$.

4. Stochastic diameter: to avoid the difficulty of calculating the stochastic distance between every pair of (s, s'), we compute its approximation by solving for the optimal value of an MDP $M_{s'}$ for each s' instead: $M_{s'}$ has the same transition function as M except that s' goes to an additional absorbing state s''; there is $+1$ reward when transitioning into s'', and 0 everywhere else. Our empirical version of T_M is then $\max_{s,s'} \log_\gamma V^{\pi^*_{M_{s'},\gamma}}_{M_{s'},\gamma}(s)$ with $\gamma = 0.9999$ (in fact, as γ tends to 1, this is equal to T_M in the limit under mild conditions).

5. ϵ-return mixing time: same as ϵ-mixing time ($\epsilon = 0.01$, checking for 50 steps).

6. Action Variation: we first compute for each $s \in S$ $\max_{a,a'\in A} \|P(\cdot|s, a) - P(\cdot|s, a')\|_1$ in the definition of δ_P. Instead of taking max over all s, we take the average to be our empirical version of δ_P.

5.3 Results

We present results for each of the following MDP distributions: `fixed(10,3)`, `binom(10,0.3)`, and `ring(10,0.125)`. For each of the 5 structural parameters we have identified, we divide the 10^5 MDPs sampled from each distribution into 3 quantiles, and plot the relative loss averaged over each quantile of MDPs in Figure 3, where rows correspond to MDP distributions and columns correspond to identified parameters.Throughout the experiments we use $\gamma_{\text{eval}} = 0.995$ and $\gamma = 0, 0.01, \ldots, 0.99$.

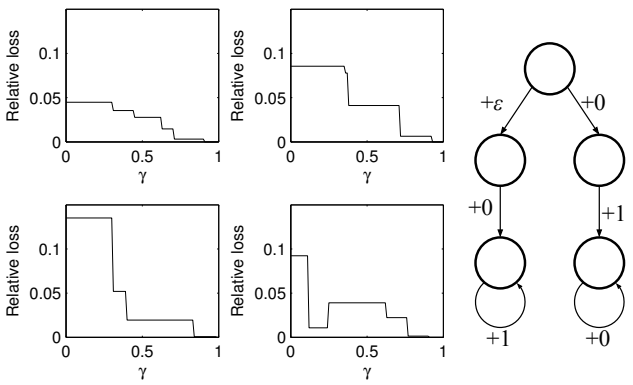


Figure 2: **Left:** relative loss as a function of γ for 4 MDPs drawn from `fixed(10,3)`. In the last graph, loss is not a monotonic function of γ. While this may be surprising, it is actually easy to construct a simple MDP where this is true (see right). **Right:** a small MDP where planning loss is non-monotonic in γ when γ_{eval} is close to 1; all transitions are deterministic and the numbers on the edges represent rewards; ϵ is a small number close to 0. The left action is optimal for γ_{eval} close to 1, and is taken when $\gamma = 0$; when $\gamma = 2\epsilon$, however, the agent will take the right action.

As seen in Figure 3, although the loss for individual MDPs are piecewise constant curves and can have complicated shapes (see Figure 2), when averaged over a distribution over MDPs we get smooth loss curves monotonically decreasing with γ. Secondly, for each parameter, the loss curves for different quantiles are separated and exhibit the order predicted by our theoretical results (except in a few cases where the separation is not significant): all our bounds are monotonically increasing with the parameters, and in the results we see the loss curves corresponding to higher quantiles stay above those for lower quantiles, which validates our theoretical results.

6 Related Work

The simple bound obtained by combining Theorem 1 and Proposition 1 is very similar to that given in [Kearns *et al.*, 2002] where a discrete horizon is used instead of a continuous discount factor, and similar results have been implied in the convergence of value iteration [Sutton and Barto, 1998]. When planning with an inaccurate model under the problem specified horizon, the dependence of loss on planning horizon is well understood, especially when the model inaccuracy is due to statistical estimation errors [Mannor *et al.*, 2007; Maillard *et al.*, 2014] or approximation errors due to the use of function approximators [Ravindran and Barto, 2004; Taylor *et al.*, 2009; Farahmand *et al.*, 2010], or the combination of the two [Paduraru *et al.*, 2008; Jiang *et al.*, 2015a].

The papers mentioned above do not address the setting

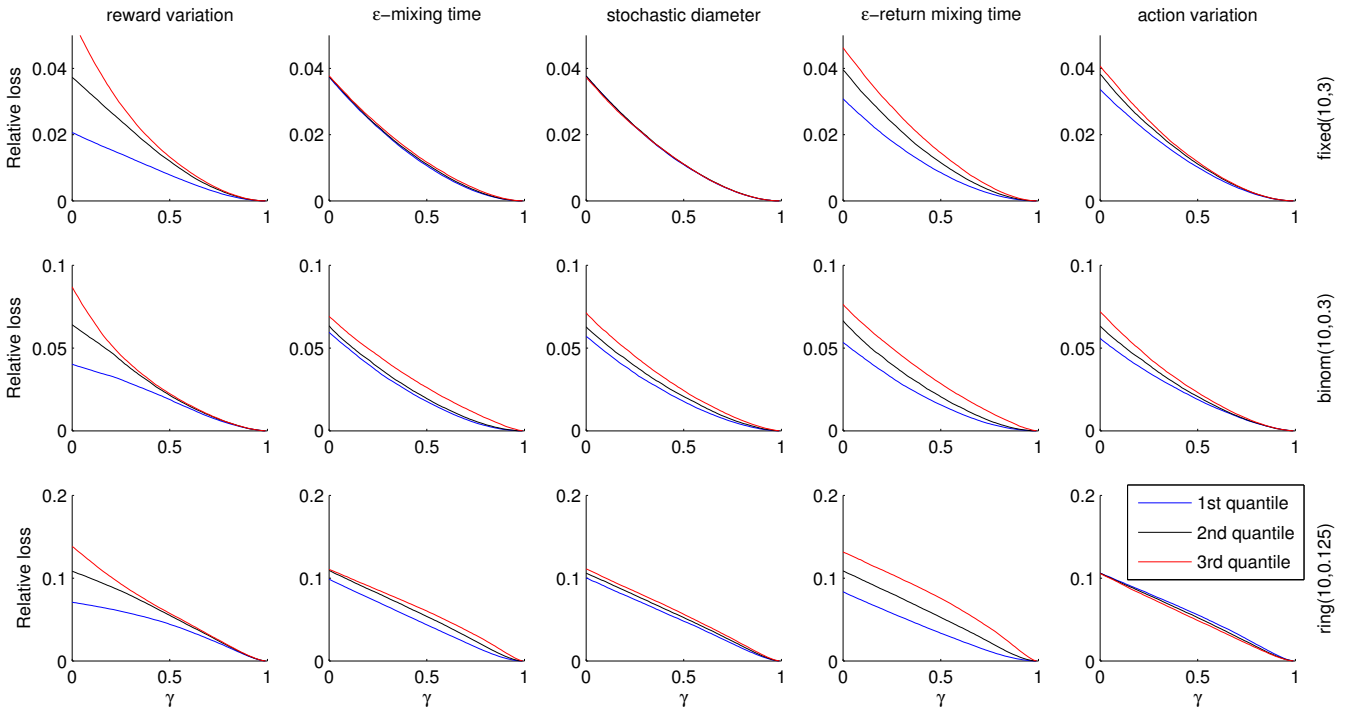


Figure 3: Experiment results on random MDPs. Each figure displays the relative loss as a function of γ, averaged over a particular distribution of MDPs (see distribution names at the right end and their descriptions in Section 5.1) in different quantiles partitioned according to a particular parameter (see parameter names at the top and their descriptions in Section 5.2). The loss curves are all well separated with an expected order, except for ϵ-mixing time and stochastic diameter with `fixed(10,3)`, and action variation with `ring(10,0.125)`.

where $\gamma < \gamma_{\text{eval}}$, and as far as we know, Petrik & Scherrer [2009] were the first to examine this setting in the particular scenario where approximation schemes are deployed in dynamic programming. More recently, Jiang *et al.* [2015b] studied another important scenario where there exists statistical estimation error in the model. In both these papers, the focus is on how the (negative) impact of model error grows as γ increases; to obtain the best planning quality, such an impact has to be traded-off with the loss incurred by using γ when planning with a perfect model, and their treatments for this loss were primary. Characterizing such a loss using structural properties of the MDP is exactly the topic of our paper, and we believe our study complements that of Petrik & Scherrer and Jiang *et al.*, and provides a more complete picture of planning with smaller than specified horizons.

The structural parameters identified in this paper are inspired by some existing work: the notion of mixing times is often used in average time reward MDPs, e.g., [Kearns and Singh, 2002; Brafman and Tennenholtz, 2003], and a term similar to our definition of stochastic diameter is defined in [Tewari and Bartlett, 2008]. As far as we know, the other two parameters (reward variation and action variation), as well as the application of all of these parameters to bounding planning loss of shallow planning, are novel.

7 Conclusions

In this paper we presented multiple high-level structural properties of MDPs that upper-bound the loss due to shallow planning with accurate models. Empirical results validated the role of these properties using a form of smoothed analysis.

Our theoretical results are also relevant to the setting of planning with inaccurate models learned from data as follows. As shown in Jiang *et al.* [2015b] an upper bound on the loss due to shallow planning with inaccurate models can be decomposed into two terms, an estimation error term that captures the loss due to the limited amount of data used to learn the model, and an approximation error term that captures the loss due to shallow planning. Our theoretical results can be viewed as providing structural parameters that affect the approximation error term.

Finally, our work provides the theoretical foundation for developing MDP planning algorithms that automatically choose an appropriate horizon. In fact, direct corollaries of our theory already offer some guidance on how to make such a choice: for example, if one has planned with a relatively small γ, he/she can read-off the variation of the resulting value function (which is κ_γ) and infer a loss bound via Theorem 2. If the loss is affordable, he/she can choose not to re-plan with a larger γ in order to save computation. There is more work to be done towards a practical algorithm, and we leave this possibility for future exploration.

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.

References

[Brafman and Tennenholtz, 2003] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*, 3:213–231, 2003.

[Farahmand *et al.*, 2010] 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.

[Hordijk and Yushkevich, 2002] Arie Hordijk and Alexander A Yushkevich. Blackwell optimality. In *Handbook of Markov decision processes*, pages 231–267. Springer, 2002.

[Jiang *et al.*, 2015a] Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction Selection in Model-based Reinforcement Learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 179–188, 2015.

[Jiang *et al.*, 2015b] Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The Dependence of Effective Planning Horizon on Model Accuracy. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems*, pages 1181–1189, 2015.

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

[Kearns *et al.*, 2002] 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.

[Levin *et al.*, 2009] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times*. American Mathematical Soc., 2009.

[Li *et al.*, 2010] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670. ACM, 2010.

[Maillard *et al.*, 2014] Odalric-Ambrym Maillard, Timothy A Mann, and Shie Mannor. "How hard is my MDP?" The distribution-norm to the rescue. In *Advances in Neural Information Processing Systems*, pages 1835–1843, 2014.

[Mannor *et al.*, 2007] 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.

[Paduraru *et al.*, 2008] Cosmin Paduraru, Robert Kaplow, Doina Precup, and Joelle Pineau. Model-based reinforcement learning with state aggregation. In *8th European Workshop on Reinforcement Learning*, 2008.

[Petrik and Scherrer, 2009] 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.

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

[Spielman and Teng, 2009] Daniel A Spielman and Shang-Hua Teng. Smoothed analysis: an attempt to explain the behavior of algorithms in practice. *Communications of the ACM*, 52(10):76–84, 2009.

[Sutton and Barto, 1998] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*. MIT Press, 1998.

[Taylor *et al.*, 2009] Jonathan Taylor, Doina Precup, and Prakash Panagaden. Bounding Performance Loss in Approximate MDP Homomorphisms. In *Advances in Neural Information Processing Systems*, pages 1649–1656, 2009.

[Tewari and Bartlett, 2008] Ambuj Tewari and Peter L Bartlett. Optimistic Linear Programming gives Logarithmic Regret for Irreducible MDPs. In *Advances in Neural Information Processing Systems*, pages 1505–1512, 2008.

A Proofs

A.1 Full proof of Theorem 1

First notice that

$$
\begin{aligned}
{\gamma_{\text{eval}}}^0 &= \gamma^0 \\
{\gamma_{\text{eval}}}^1 &= \gamma^1 + (\gamma_{\text{eval}} - \gamma) \\
{\gamma_{\text{eval}}}^2 &= \gamma^2 + \gamma(\gamma_{\text{eval}} - \gamma) + \gamma_{\text{eval}}(\gamma_{\text{eval}} - \gamma) \\
&\vdots \\
{\gamma_{\text{eval}}}^k &= \gamma^k + \gamma^{k-1}(\gamma_{\text{eval}} - \gamma) + \gamma^{k-2}\gamma_{\text{eval}}(\gamma_{\text{eval}} - \gamma) + \cdots \\
&\quad + {\gamma_{\text{eval}}}^{k-1}(\gamma_{\text{eval}} - \gamma) \\
&\vdots
\end{aligned}
$$

On the right-hand side of the equations, each column of the array forms a geometric series with ratio γ. This means that for any policy π, we can decompose $V^\pi_{\gamma_{\text{eval}}}$ into a linear combination of V^π_γ by decomposing and re-arranging the reward obtained at each step. In particular, $\forall s \in S$, let e_s be the unit vector with the element indexed by s equal to 1, then

$$
\begin{aligned}
&V^\pi_{\gamma_{\text{eval}}}(s) \\
&= e_s^\top \sum_{t=1}^{\infty} {\gamma_{\text{eval}}}^{t-1}[P^\pi]^{t-1}R^\pi \\
&= e_s^\top \sum_{t=1}^{\infty} \gamma^{t-1}[P^\pi]^{t-1}R^\pi + (\gamma_{\text{eval}} - \gamma)e_s^\top \sum_{t=1}^{\infty} \gamma^{t-1}[P^\pi]^{t}R^\pi + \\
&\quad \cdots + (\gamma_{\text{eval}} - \gamma){\gamma_{\text{eval}}}^{k-1}e_s^\top \sum_{t=1}^{\infty} \gamma^{t-1}[P^\pi]^{t+k-1}R^\pi + \cdots \\
&= e_s^\top V^\pi_\gamma + (\gamma_{\text{eval}} - \gamma)e_s^\top [P^\pi]V^\pi_\gamma + \cdots \\
&\quad + (\gamma_{\text{eval}} - \gamma){\gamma_{\text{eval}}}^{k-1}e_s^\top [P^\pi]^k V^\pi_\gamma + \cdots,
\end{aligned}
$$

where $[P^\pi]$ is a $|S| \times |S|$ matrix with the element indexed by (s, s') being $P(s'|s, a)$, and R^π is a $|S| \times 1$ vector with the s-th element being $R(s, \pi(s))$. Now plug in π^*_γ and $\pi^*_{\gamma_{\text{eval}}}$ and consider the value difference: $\forall s \in S$,

$$
\begin{aligned}
&V^{\pi^*_{\gamma_{\text{eval}}}}_{\gamma_{\text{eval}}}(s) - V^{\pi^*_\gamma}_{\gamma_{\text{eval}}}(s) \\
&= \big(e_s^\top V^{\pi^*_{\gamma_{\text{eval}}}}_\gamma - e_s^\top V^{\pi^*_\gamma}_\gamma\big) + \\
&\quad (\gamma_{\text{eval}} - \gamma)\big(e_s^\top [P^{\pi^*_{\gamma_{\text{eval}}}}]V^{\pi^*_{\gamma_{\text{eval}}}}_\gamma - e_s^\top [P^{\pi^*_\gamma}]V^{\pi^*_\gamma}_\gamma\big) + \cdots + \\
&\quad (\gamma_{\text{eval}} - \gamma){\gamma_{\text{eval}}}^{k-1}\big(e_s^\top [P^{\pi^*_{\gamma_{\text{eval}}}}]^k V^{\pi^*_{\gamma_{\text{eval}}}}_\gamma - e_s^\top [P^{\pi^*_\gamma}]^k V^{\pi^*_\gamma}_\gamma\big) \\
&\quad + \cdots \\
&\le \big(e_s^\top V^{\pi^*_\gamma}_\gamma - e_s^\top V^{\pi^*_\gamma}_\gamma\big) + \\
&\quad (\gamma_{\text{eval}} - \gamma)\big(e_s^\top [P^{\pi^*_{\gamma_{\text{eval}}}}]V^{\pi^*_\gamma}_\gamma - e_s^\top [P^{\pi^*_\gamma}]V^{\pi^*_\gamma}_\gamma\big) + \cdots + \\
&\quad (\gamma_{\text{eval}} - \gamma){\gamma_{\text{eval}}}^{k-1}\big(e_s^\top [P^{\pi^*_{\gamma_{\text{eval}}}}]^k V^{\pi^*_\gamma}_\gamma - e_s^\top [P^{\pi^*_\gamma}]^k V^{\pi^*_\gamma}_\gamma\big) \\
&\quad + \cdots \\
&\le 0 + (\gamma_{\text{eval}} - \gamma)\kappa_\gamma + \cdots + (\gamma_{\text{eval}} - \gamma){\gamma_{\text{eval}}}^{k-1}\kappa_\gamma + \cdots \\
&= \frac{\gamma_{\text{eval}} - \gamma}{1 - \gamma_{\text{eval}}}\kappa_\gamma.
\end{aligned}
$$

The first inequality above holds by optimality of π^*_γ for discount factor γ and the second inequality holds because for any stochastic vectors p, q, we have $|p^\top V^{\pi^*_\gamma}_\gamma - q^\top V^{\pi^*_\gamma}_\gamma| \le \kappa_\gamma$. □

A.2 Proof of Claim 1

Prove by construction: consider a two-state MDP where s_2 is absorbing with reward $R_{\max}$; s_1 has two actions, with a_1 absorbing giving reward $R(s_1, a_1) = r$ (r is a real number to be set later), and a_2 transitioning to s_2 with reward $R(s_1, a_2) = (r - \gamma R_{\max})/(1 - \gamma)$. It is easy to verify that $R(s_1, a_2) \le R(s_1, a_1)$, hence $\Delta_R = R_{\max} - r$. We check the loss at s_1: the MDP is designed so that $\pi^*_{\gamma_{\text{eval}}}(s_1) = a_2$ and $\pi^*_\gamma(s_1) = a_1$, and the loss of using γ is $\frac{(R_{\max} - r)(\gamma_{\text{eval}} - \gamma)}{(1-\gamma)(1-\gamma_{\text{eval}})}$, which is exactly equal to the bound since $\frac{R_{\max} - r}{1-\gamma} = \Delta_R/(1 - \gamma) = \kappa_\gamma$. □

A.3 Proof of Proposition 3

We reduce Proposition 3 to Proposition 5 (presented below): if π has ϵ-mixing time $T_\pi(\epsilon)$ w.r.t. ρ^π, we have for any $t \ge T_\pi(\epsilon), s \in S$,

$$
\begin{aligned}
&\big|e_s^\top [P^\pi]^t R^\pi - \eta^\pi\big| = \big|e_s^\top [P^\pi]^t R^\pi - (\rho^\pi)^\top R^\pi\big| \\
&\le \big\|e_s^\top [P^\pi]^t - (\rho^\pi)^\top\big\|_1 R_{\max}/2 = \epsilon R_{\max}/2.
\end{aligned}
$$

Therefore $T_\pi(\epsilon)$ is also an $(\epsilon R_{\max}/2)$-return mixing time for π, and applying Proposition 5 the result follows. □

A.4 Proof of Proposition 5

Let $\overline{V}^\pi_\gamma = V^\pi_\gamma - \frac{\gamma^{T^v_\pi(\epsilon)}}{1-\gamma}\eta^\pi$, which is the value function offset by a state-independent constant. For any $s, s' \in S$,

$$
V^\pi_\gamma(s) - V^\pi_\gamma(s') = \overline{V}^\pi_\gamma(s) - \overline{V}^\pi_\gamma(s'),
$$

and $\overline{V}^\pi_\gamma$ is equal to

$$
\begin{aligned}
&\sum_{t=1}^{\infty} \gamma^{t-1}[P^\pi]^{t-1}R^\pi - \frac{\gamma^{T^v_\pi(\epsilon)}}{1-\gamma}\eta^\pi \\
&= \sum_{t=1}^{T^v_\pi(\epsilon)} \gamma^{t-1}[P^\pi]^{t-1}R^\pi + \sum_{t=T^v_\pi(\epsilon)+1}^{\infty} \gamma^{t-1}[P^\pi]^{t-1}R^\pi - \frac{\gamma^{T^v_\pi(\epsilon)}}{1-\gamma}\eta^\pi \\
&= \sum_{t=1}^{T^v_\pi(\epsilon)} \gamma^{t-1}[P^\pi]^{t-1}R^\pi + \sum_{t=T^v_\pi(\epsilon)+1}^{\infty} \gamma^{t-1}\Big([P^\pi]^{t-1}R^\pi - \eta^\pi\Big).
\end{aligned}
$$

We now have,

$$
\begin{aligned}
&\overline{V}^\pi_\gamma(s) - \overline{V}^\pi_\gamma(s') \\
&\le \sum_{t=1}^{T^v_\pi(\epsilon)} \gamma^{t-1}(e_s - e_{s'})^\top [P^\pi]^{t-1}R^\pi \\
&\quad + 2 \cdot \left\| \sum_{t=T^v_\pi(\epsilon)+1}^{\infty} \gamma^{t-1}\big([P^\pi]^{t-1}R^\pi - \eta^\pi\big) \right\|_\infty \\
&\le \frac{1 - \gamma^{T^v_\pi(\epsilon)}}{1-\gamma}R_{\max} + 2 \sum_{t=T^v_\pi(\epsilon)+1}^{\infty} \gamma^{t-1}\epsilon \\
&= \frac{R_{\max}(1 - \gamma^{T^v_\pi(\epsilon)}) + 2\epsilon\gamma^{T^v_\pi(\epsilon)}}{1-\gamma}.
\end{aligned}
$$

□

A.5 Proof of Theorem 2

We first prove a helping lemma widely used in MDP approximation literature.

Lemma 2. *Given stochastic vectors $p, q \in \mathbb{R}^{|S|}$, and a real vector V with the same dimension,*

$$|p^\top V - q^\top V| \le \|p - q\|_1 \max_{s,s'} |V(s) - V(s')|/2.$$

Proof. Let $c = (\max_s V(s) + \min_s V(s))/2$,

$$\begin{aligned} |p^\top V - q^\top V| &= |p^\top (V - c) - q^\top (V - c)| \\ &\le \|p - q\|_1 \|V - c\|_\infty \text{(Hölder's inequality)} \\ &= \|p - q\|_1 \max_{s,s'} |V(s) - V(s')|/2. \end{aligned}$$ □

Proof of Theorem 2. The proof follows from the proof for Theorem 1 up to Equation 3. The k-th term in the summation is (ignoring a factor of $(\gamma_{\text{eval}} - \gamma){\gamma_{\text{eval}}}^{k-1}$ for the moment):

$$\begin{aligned} & \left(e_s^\top [P^{\pi^*_{\gamma_{\text{eval}}}}]^k V_\gamma^{\pi^*_\gamma} - e_s^\top [P^{\pi^*_\gamma}]^k V_\gamma^{\pi^*_\gamma}\right) \\ \le & \left\| e_s^\top [P^{\pi^*_{\gamma_{\text{eval}}}}]^k - e_s^\top [P^{\pi^*_\gamma}]^k \right\|_1 \kappa_\gamma/2 \qquad \text{(Lemma 2)} \\ \le & \left(2 - 2(1 - \delta_P/2)^k\right)\kappa_\gamma/2. \qquad \text{(Corollary 4)} \end{aligned}$$

Summing over $k = 1, 2, \ldots$, we have the loss bounded by

$$\begin{aligned} & \textstyle\sum_{k=1}^\infty (\gamma_{\text{eval}} - \gamma){\gamma_{\text{eval}}}^{k-1}\left(1 - (1 - \delta_P/2)^k\right)\kappa_\gamma \\ = & \frac{\gamma_{\text{eval}} - \gamma}{1 - \gamma_{\text{eval}}}\kappa_\gamma - \frac{(\gamma_{\text{eval}} - \gamma)(1 - \delta_P/2)}{1 - \gamma_{\text{eval}}(1 - \delta_P/2)}\kappa_\gamma, \end{aligned}$$

which is the RHS of the bound after simplification. □

A.6 Proof of Claim 2

We construct the following 3-state MDP (see Figure 1b): s_2 and s_3 are absorbing with rewards $R_{\max}$ and 0 respectively; consequently $\kappa_\gamma = R_{\max}/(1 - \gamma)$. For s_1, there are two actions a_1 and a_2, with reward and transition rules as follows: given a real number $c \in [0, 1]$ to be set later, we set $R(s_1, a_1) = 0$, $R(s_1, a_2) = \gamma(1 - c)R_{\max}/(1 - \gamma)$, and $P(s_1|s_1, a_1) = P(s_1|s_1, a_2) = c$, $P(s_2|s_1, a_1) = P(s_3|s_1, a_2) = 1 - c$. In this MDP, we have $\delta_P = 2 - 2c$, so we can manipulate δ_P by setting $c = 1 - \delta_P/2$; also, both actions in s_1 are equally good under γ but $\pi^*_{\gamma_{\text{eval}}}(s_1) = a_1$, and the loss is

$$\begin{aligned} & \frac{\gamma_{\text{eval}}(1 - c)R_{\max}}{(1 - \gamma_{\text{eval}})(1 - c\gamma_{\text{eval}})} - \frac{\gamma(1 - c)R_{\max}}{(1 - \gamma)(1 - c\gamma_{\text{eval}})} \\ = & \frac{(1 - c)(\gamma_{\text{eval}} - \gamma)R_{\max}}{(1 - \gamma_{\text{eval}})(1 - \gamma_{\text{eval}}c)(1 - \gamma)} \\ = & \frac{\delta_P/2(\gamma_{\text{eval}} - \gamma)\kappa_\gamma}{(1 - \gamma_{\text{eval}})(1 - \gamma_{\text{eval}}(1 - \delta_P/2))}, \end{aligned}$$

which is just the RHS of the bound. □

B Empirical results including the distributions of the identified parameters

In this section we present the results shown in Figure 3 together with the distribution of the identified parameters. The left columns of Figure 4, 5, and 6 are exactly the rows of Figure 3, and the right columns show the distribution of each parameter in each of these MDP distributions.

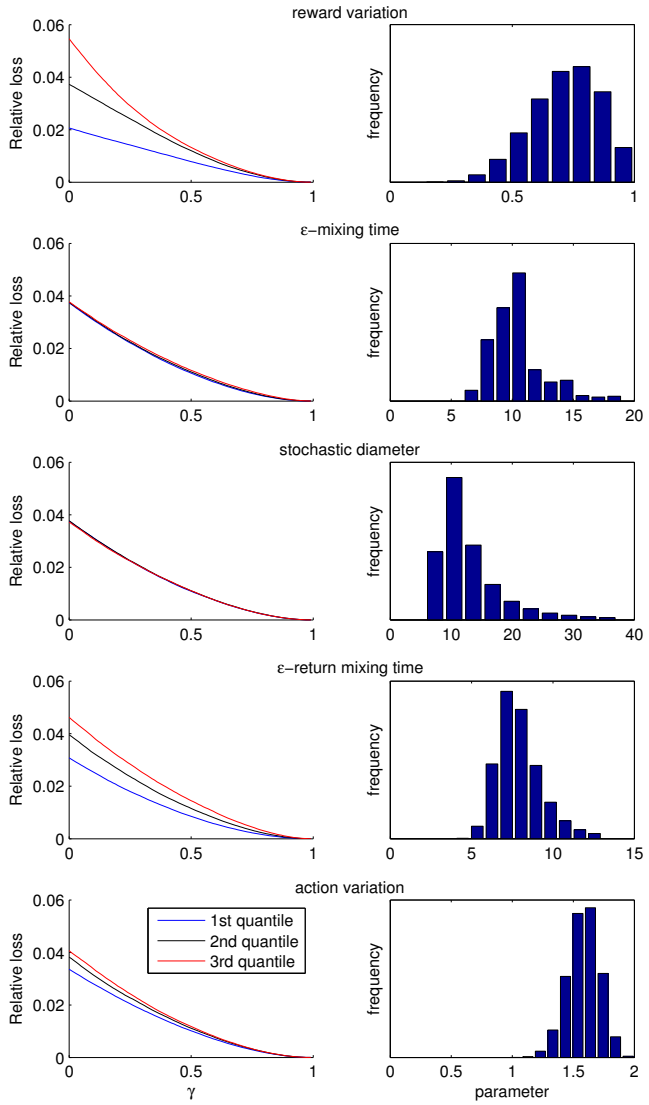


Figure 4: Results on random MDPs drawn from `fixed(10,3)`. The left column shows the relative loss averaged over MDPs in different quantiles partitioned according to each of the parameters (see the list in Section 5.2), and the right column shows the distribution of the parameters. For this distribution of MDPs (`fixed(10,3)`), the loss curves are well separated with an expected order for all the parameters except ϵ-mixing time and stochastic diameter.

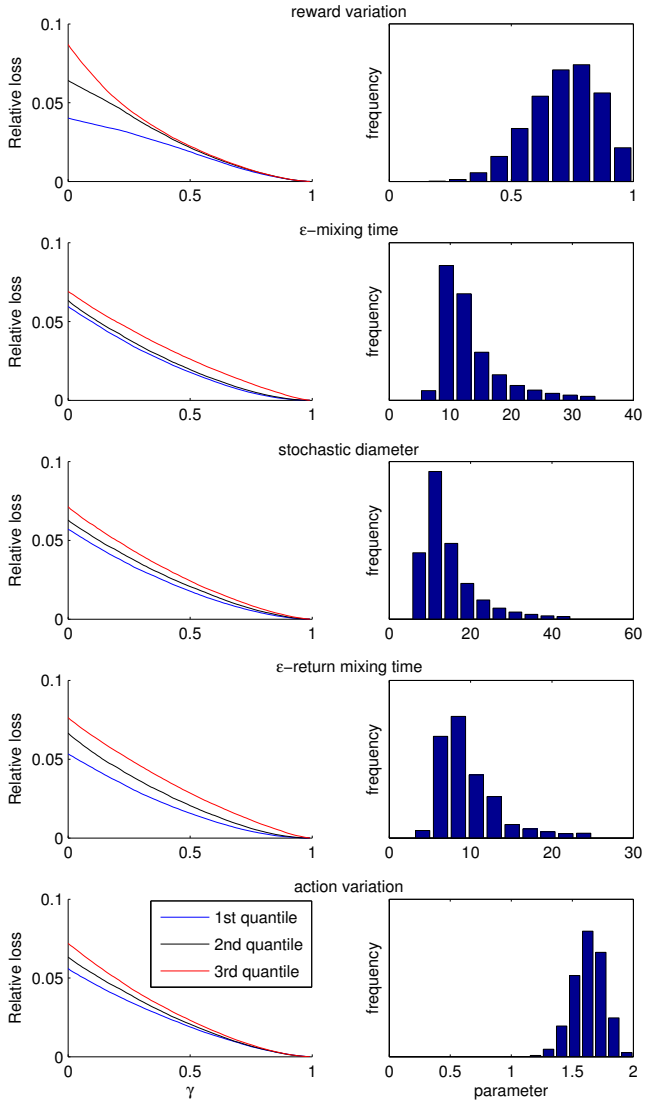


Figure 5: Results on random MDPs drawn from `binom(10,0.3)`. The left column shows the relative loss averaged over MDPs in different quantiles partitioned according to each of the parameters, and the right column shows the distribution of the parameters. For this distribution of MDPs (`binom(10,0.3)`), the curves are all well separated with an expected order.

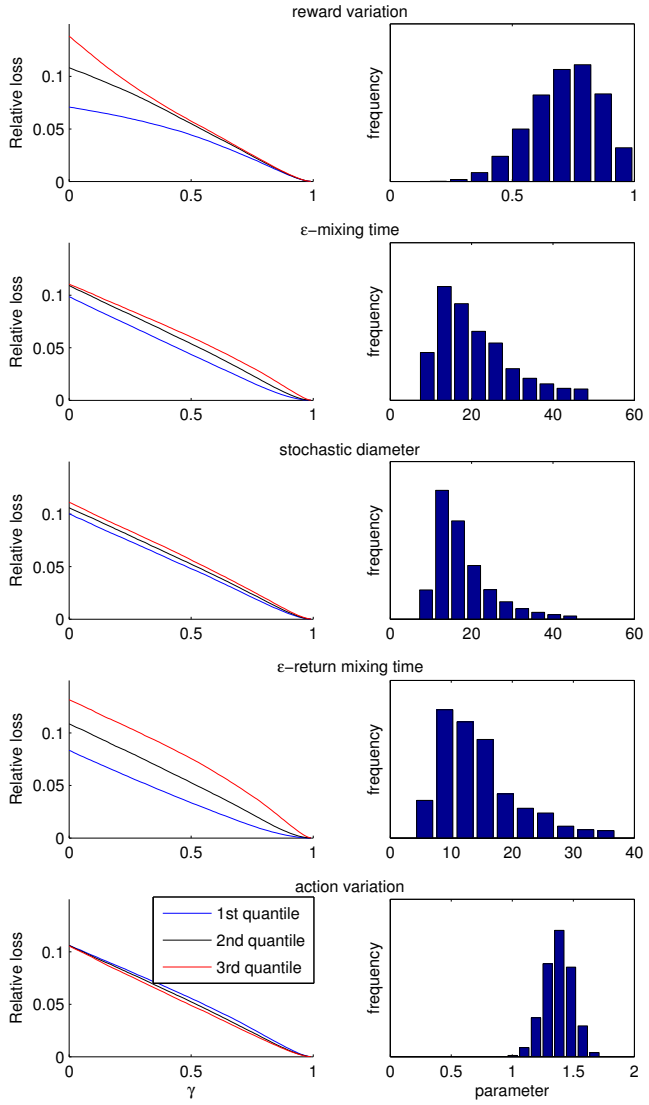


Figure 6: Results on random MDPs drawn from `ring(10,0.125)`. The left column shows the relative loss averaged over MDPs in different quantiles partitioned according to each of the parameters, and the right column shows the distribution of the parameters. For this distribution of MDPs (`ring(10,0.125)`), the curves are all well separated with an expected order except for action variation.