On shallow planning under partial observability

Randy Lefebvre
randy.lefebvre.1@ulaval.ca
Université Laval

Audrey Durand
audrey.durand@ift.ulaval.ca
Université Laval
Canada CIFAR AI Chair

Abstract

Formulating a real-world problem under the Reinforcement Learning framework involves non-trivial design choices, such as selecting a discount factor for the learning objective (discounted cumulative rewards), which articulates the planning horizon of the agent. This work investigates the impact of the discount factor on the bias-variance trade-off given structural parameters of the underlying Markov Decision Process. Our results support the idea that a shorter planning horizon might be beneficial, especially under partial observability.

1 Introduction

Reinforcement Learning (RL) has had tremendous success on Atari games (Mnih et al., 2013), yet applications of RL in the real-world remain limited (Dulac-Arnold et al., 2021). This complexity is due to many challenges such as sample efficiency of RL methods (Yu, 2018), risk/safety issues (Gu et al., 2023) and partial observability (Sondik, 1978; Francois-Lavet et al., 2019; Kaelbling et al., 1998). Formulating a real-world problem under the RL framework also involves several non-trivial decisions such as selecting a state/action space (especially when these are continuous), formulating a reward function, and formulating the learning objective (Hare, 2019; Devidze et al., 2021). The learning objective usually corresponds to the discounted cumulative rewards, which depends on a discount factor articulating the considered planning horizon when attributing values to states and actions (Sutton and Barto, 2018). This objective is useful since it can reduce the search space intuitively by giving less credit to futur rewards and actions. In toy and simulated environments, early practitioners tend to use large discount factor values often found in the RL literature on Atari (Kaiser et al., 2024; Mnih et al., 2013). This equates to considering very long planning horizons. On the other hand, real-world applications tend to formulate sequential decision-making problems under the contextual bandit setting (i.e. with a myopic agent w.r.t. the planning horizon) in response to low data regimes (Bastani and Bayati, 2020; Ding et al., 2019; Durand et al., 2018).

The impact of reducing the planning horizon has been studied previously, and bounds on the resulting bias-variance trade-off on the state value functions have been proposed (Amit et al., 2020; Jiang et al., 2015). Unfortunately, these results provide loose bounds that do not consider the structure of the underlying Markov Decision Process (MDP) and thus fail to capture its impact on the *optimal planning horizon*. The optimal planning horizon can be described as the discount factor γ which minimizes the planning loss (see Eq. 1) i.e. the one which can extract the best policy possible considering the limited amount of data. Distinct results involving the structure of MDPs (Jiang et al., 2016; Gheshlaghi Azar et al., 2013; Wu et al., 2023; He et al., 2021) have been achieved separately, but these insights have never been brought together.

Contributions In this work, we introduce new results on the bias-variance trade-off that explicitly depend on high-level structural parameters of the underlying MDP (Section 2). More importantly, our results touch on Partially Observable MDPs (POMDPs), providing the first insights supporting the advantage of considering short horizons in the learning objective for practical applications under partial observability (Section 3). We support and illustrate the theory with numerical results

(Section 4), hoping that this can open the door to choices in learning objectives that are better adapted to real-world RL applications. Finally, we make the code to recreate our simulations and results open source to ensure reproducibility and offer a framework which practitioners can modify to better understand the impact of partial observability on their specific applications.

2 Fully observable setting

An MDP can be described as a tuple $(\mathcal{S}, \mathcal{A}, P, R)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is a transition function, and $R : \mathcal{S} \times \mathcal{A} \mapsto [0, R_{\max}]$ is a reward function, with $R_{\max}$ denoting the maximal reward obtainable in the MDP. On each time step $t \in \mathbb{N}_0$, the current state $S_t \in \mathcal{S}$ is observed, an action $A_t \in \mathcal{A}$ is played, the environment transitions into next state S_{t+1} (using P) and generates an observed reward R_{t+1} (using R). Given an MDP (tuple) M, the value of state $s \in \mathcal{S}$ under policy $\pi : \mathcal{S} \mapsto \mathcal{A}$ is the expected sum of discounted rewards obtained by selecting actions according to policy π from state s:

$$V^{\pi}_{M,\gamma}(s) = \mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s\right],$$

where the discount factor $\gamma \in [0, 1]$ controls the effective planning horizon (the credit assigned to action A_t for future rewards). The goal of a learning agent is to find the optimal policy $\pi^*_{M,\gamma}$ that maximizes $V^{\pi}_{M,\gamma}(s)$ for all states $s \in \mathcal{S}$. We use $V^{\pi}_{M,\gamma} \in \mathbb{R}^{|\mathcal{S}|}$ to denote the vector of state values. A full table containing all the notation in the paper can be found in Appendix A.

Blackwell discount factor Practitioners often believe using a higher discount factor will get a better policy on their specific problem. While this is true with an infinite amount of data, it is rarely the case when building RL applications in practice. It has even been shown previously that there always exists a discount factor in finite MDPs such that the optimal policy cannot be improved by further extending the planning horizon under the Blackwell optimality criterion when $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$ (Grand-Clément and Petrik (2023) Thm. 3.2). Above this point, we are in fact only cumulating variance and noise. We refer to the corresponding discount factor as the *Blackwell discount factor* denoted γ_{Bw}. We refer to the planning horizon under discount factor γ_{Bw} as the *Blackwell planning horizon.* This concept closely resembles the idea of *Effective planning horizon* in Laidlaw et al. (2023), but with a discount factor instead of a number of steps look ahead.

Planning loss The planning loss captures the impact of using $\gamma < \gamma_{Bw}$ given that the planning is performed in an approximate model of the environment $\widehat{M}$ with $\widehat{M} \approx M$[1]:

$$\begin{aligned} \|V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{Bw}}\|_\infty &= \|V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}} - V^{\pi^*_{M,\gamma}}_{M,\gamma_{Bw}} + V^{\pi^*_{M,\gamma}}_{M,\gamma_{Bw}} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{Bw}}\|_\infty \\ &\le \underbrace{\|V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}} - V^{\pi^*_{M,\gamma}}_{M,\gamma_{Bw}}\|_\infty}_{\text{bias}} + \underbrace{\|V^{\pi^*_{M,\gamma}}_{M,\gamma_{\text{Bw}}} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma_{Bw}}\|_\infty}_{\text{variance}}. \end{aligned} \tag{1}$$

This decomposition offers insight into two components which can affect the quality of the policy obtained when planning on an approximate model of the environment using a shallow planning horizon ($\gamma < \gamma_{Bw}$). The *bias* denotes the loss in value function (evaluated on the true MDP M and with the Blackwell planning horizon) when using a policy that is optimal with a shallow planning horizon γ instead of using a policy that is optimal with γ_{Bw}. On the other hand, the *variance* captures the impact of optimizing the policy under an approximate model $\widehat{M}$ with a shallow planning horizon and will tend to 0 with more data. This decomposition is different from previous work (Jiang et al., 2015) and has the advantage of being interpretable as a bias-variance trade-off from the PAC-learning literature. We can compare the bias to the approximation error and the variance to the estimation error (Shalev-Shwartz and Ben-David, 2014).

[1] We remain agnostic to how $\widehat{M}$ is computed and how the data is collected as literature on the topic is abundant (Gheshlaghi Azar et al., 2013; Wu et al., 2023; He et al., 2021).

Structural parameters have been introduced previously to characterize the difficulty of an MDP.

Definition 1 (Value-function variation, Jiang et al. (2016))**.** *Given an MDP M and discount factor γ, the value-function variation captures the largest difference between the values of two different states when following the optimal policy:*

$$\kappa_{M,\gamma} = \max_{s,s' \in \mathcal{S}} \left| V_{M,\gamma}^{\pi^*_{M,\gamma}}(s) - V_{M,\gamma}^{\pi^*_{M,\gamma}}(s') \right| \leq \frac{R_{max}}{1-\gamma}.$$

Note that the value-function is evaluated with the same discount horizon as used by the policy.

The value-function variation provides insight on the impact of starting in certain states over others in the MDP. A low value indicates that the starting state is not important to consider for predicting future rewards (Jiang et al., 2016).

Definition 2 (Action variation, Jiang et al. (2016))**.** *Given an MDP M with transition probabilities P, the action variation captures how much actions can impact state transitions:*

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

If the action variation is equal to 0, the agent cannot influence the state transitions (and therefore future rewards). In this case, we would expect the problem to be safely (and efficiently) formulated under the contextual bandit setting, which corresponds to using a myopic ($\gamma = 0$) agent. The maximal value for this L_1 distance is 2, which often happen under deterministic settings.

Using Definitions 1 and 2, Jiang et al. (2016) introduced the following bound on the bias:

$$||\underbrace{V_{M,\gamma_{Bw}}^{\pi^*_{M,\gamma_{Bw}}} - V_{M,\gamma_{Bw}}^{\pi^*_{M,\gamma}}}_{\text{bias}}||_\infty \leq \frac{\delta_M/2 \cdot \kappa_{M,\gamma}(\gamma_{Bw} - \gamma)}{(1-\gamma_{Bw})(1-\gamma_{Bw}(1-\delta_M/2))}. \tag{2}$$

Unfortunately, the action variation is not sensitive to the planning horizon of the agent compared with the Blackwell planning horizon. Moreover, unlike the bias, there is no current bounds for the variance. We address these limitations in Sections 2.1 and 2.2, which will allow us to obtain a new bound on the planning loss in Section 2.3.

2.1 Improving the bias bound

In order to consider the planning horizon in the bias bound of Jiang et al. (2016) (Eq. 2), we introduce the following definitions:

Definition 3 (Discordant state-action pairs)**.** *The set of state-action pairs in an MDP M where two policies π and π' differ:*

$$\mathcal{Z}_M(\pi \neq \pi') = \{(s,a) \in \mathcal{S} \times \mathcal{A} : \pi(s) \neq \pi'(s),\ \pi'(s) = a\}.$$

This new set will be used to capture the impact of a shallow-horizon policy on the action variation:

Definition 4 (Horizon-sensitive action variation)**.** *The most important difference in transition probabilities induced by using discount factor γ instead of discount factor γ_{Bw} on an MDP M:*

$$\delta_{M,\gamma} = \max_{(s,a) \in \mathcal{Z}_M(\pi^*_{M,\gamma} \neq \pi^*_{M,\gamma_{Bw}})} \|P(\cdot|s,\pi^*_{M,\gamma}(s)) - P(\cdot|s,a)\|_1.$$

The implementation of the action variations in proofs is to bound the difference in transition probabilities between two different policies. The highest possible bound is given by prior results (Definition 2), but we tighten this result by simply considering states for which the policies are unequal instead of all states. This has the benefit of being 0 when the policy is evaluated with a discount factor above the Blackwell. By building on the analysis of Jiang et al. (2016) we can obtain the following result to characterize the impact of optimizing the policy with a shallow horizon on k-steps transition probabilities.

Proposition 1 (Horizon-sensitive transition probabilities distance)**.** *Given an MDP M, let $P^{\pi}_{s,k}$ denote the vector of the transition probabilities from state $s \in \mathcal{S}$ to every possible states when following policy π for $k \geq 1$ time steps. The transition probabilities when following the policy that is optimal for a shallow planning horizon (γ) instead of following the policy that is optimal for the Blackwell planning horizon is bounded by:*

$$\|P^{\pi^*_{M,\gamma}}_{s,k} - P^{\pi^*_{M,\gamma_{Bw}}}_{s,k}\|_1 \leq 2 - 2(1 - \delta_{M,\gamma}/2)^k.$$

Proposition 1 can then be used to extend Eq. 2 using the horizon-sensitive structural parameters:

$$\underbrace{||V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}} - V^{\pi^*_{M,\gamma}}_{M,\gamma_{Bw}}||_\infty}_{\text{bias}} \leq \frac{\delta_{M,\gamma}/2 \cdot \kappa_{M,\gamma}(\gamma_{Bw} - \gamma)}{(1 - \gamma_{Bw})(1 - \gamma_{Bw}(1 - \delta_{M,\gamma}/2))}, \tag{3}$$

where we use the term $\delta_{M,\gamma}$ (instead of δ_M) to capture the divergence between the shallow and optimal-horizon policies. Since the set offered by Def. 3 is smaller than the set of all state-action pairs, Eq. 3 is tighter than Eq. 2. See Appendix B for the complete proofs.

2.2 Controlling the variance

We will now introduce new definitions and results to provide a bound on the variance (in Eq. 1). Recall that the variance captures the impact of learning an optimal policy on an (empirical) approximation $\widehat{M}$ of a true MDP M when using a shallow planning horizon ($\gamma < \gamma_{Bw}$). To this end, it will be convenient to isolate the variance that does not depend on the shallow planning.

Definition 5 (Variance due to model approximation)**.** *The maximum difference in value-function due to the approximate model $\widehat{M}$:*

$$\hat{\epsilon} = \|V^{\pi^*_{M,\gamma}}_{M,\gamma} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma}\|_\infty.$$

This term can be upper-bounded into the following results by using known settings in the PAC literature (Gheshlaghi Azar et al., 2013; Wu et al., 2023; He et al., 2021). We can also use the discordant state-action pairs (Def. 3) to capture the action variation resulting from having optimized the policy on an approximation $\widehat{M}$ of a true MDP M.

Definition 6 (Empirical action variation)**.** *The most important difference in transition probabilities when following the policy optimal on an MDP M vs the policy optimal on an approximate model $\widehat{M}$:*

$$\hat{\delta}_{M,\gamma} = \max_{(s,a) \in \mathcal{Z}_M(\pi^*_{M,\gamma} \neq \pi^*_{\widehat{M},\gamma})} \left\| P(\cdot|s, \pi^*_{M,\gamma}(s)) - P(\cdot|s,a) \right\|_1.$$

This improvement over the action variation (Definition 2) is that it will tend towards 0 as $\widehat{M} \approx M$ which is desirable in a bound on the variance. As was done previously in Section 2.1, we can also build on the analysis of Jiang et al. (2016) to obtain the following result to characterize the impact of optimizing the policy with an approximate model $\widehat{M}$ on k-steps transition probabilities.

Proposition 2 (Empirical transition probabilities distance)**.** *Given an MDP M and an approximate model $\widehat{M}$. Let $P^{\pi}_{s,k}$ denote the vector of the transition probabilities from state $s \in \mathcal{S}$ to every possible states when following policy π for $k \geq 1$ time steps. For a planning horizon γ, the transition probabilities when following the policy that is optimal on $\widehat{M}$ instead of following the policy that is optimal on M is bounded by:*

$$\|P^{\pi^*_{M,\gamma}}_{s,k} - P^{\pi^*_{\widehat{M},\gamma}}_{s,k}\|_1 \leq 2 - 2(1 - \hat{\delta}_{M,\gamma}/2)^k.$$

Proposition 2 can be used with Definitions 3 and 6 to obtain the following bound on the variance. See Appendix C for the complete proofs.

Lemma 1 (Variance)**.** *Consider optimal policies computed with planning horizon $\gamma < \gamma_{Bw}$ on an MDP M and an approximate model $\widehat{M}$. The difference between their value-function evaluated on M with discount factor γ_{Bw} is bounded by:*

$$\|V_{M,\gamma_{Bw}}^{\pi^*_{M,\gamma}} - V_{M,\gamma_{Bw}}^{\pi^*_{\widehat{M},\gamma}}\|_\infty \leq \hat{\epsilon}\left(\frac{1-\gamma}{1-\gamma_{Bw}}\right) + \frac{\hat{\delta}_{M,\gamma}/2 \cdot \kappa_{M,\gamma}(\gamma_{Bw}-\gamma)}{(1-\gamma_{Bw})(1-\gamma_{Bw}(1-\hat{\delta}_{M,\gamma}/2))}.$$

This bound is interesting because it becomes tighter when the empirical action variation $\hat{\delta}_{M,\gamma}$ or the value function variation $\kappa_{M,\gamma}$ decrease. We can then deduct that a problem with low value in these structural parameters lowers both the bias (Eq. 3) and the variance. Finally, the use of the empirical action variation (Def. 6) gives rise to a bound that is coherent in convergence, as it will tend towards 0 as the amount of data increases.

2.3 A new bound on the planning loss

By combining the extended bias bound (Eq. 3) with our novel bound on the variance (Lemma 1), we obtain the following bound on the planning loss (Eq. 1). See Appendix D for the complete proof.

Theorem 1 (Planning loss)**.** *Given an MDP M, its Blackwell discount factor γ_{Bw}, and an approximate model $\widehat{M}$. The planning loss is bounded by:*

$$\begin{aligned}\|V_{M,\gamma_{Bw}}^{\pi^*_{M,\gamma_{Bw}}} - V_{M,\gamma_{Bw}}^{\pi^*_{\widehat{M},\gamma}}\|_\infty \leq & \kappa_{M,\gamma}\left(\frac{\gamma_{Bw}-\gamma}{1-\gamma_{Bw}}\right)\left(\frac{\delta_{M,\gamma}/2}{1-\gamma_{Bw}(1-\delta_{M,\gamma}/2)} + \frac{\hat{\delta}_{M,\gamma}/2}{1-\gamma_{Bw}(1-\hat{\delta}_{M,\gamma}/2)}\right) \\ & + \hat{\epsilon}\left(\frac{1-\gamma}{1-\gamma_{Bw}}\right).\end{aligned}$$

This result provides insight into how structural parameters affect not only the bias, but also the variance. For instance, a problem with action variation $\delta_M \approx 0$ has low variance due to the limited impact of the policy over the state value (agent actions do not impact transition probabilities). Similarly to prior work (Jiang et al., 2015), although the applicability of this result is limited by not having access to the true model M, it remains a helpful guide to design heuristics and better understand how one could decide a discount factor. For example, it justifies framing recommender systems as contextual bandits when the outcome of future recommendations do not depend on current recommendations, which translates into a low value-function variation $\kappa_{M,\gamma}$. Thm. 1 is tighter than the current existing bound (Jiang et al., 2015) under the following condition on the quality of the model approximation $\widehat{M}$ (see Appendix E):

$$\hat{\epsilon} \leq \frac{R_{max}}{1-\gamma} - \kappa_{M,\gamma}\left(\frac{\delta_{M,\gamma}/2}{1-\gamma_{Bw}(1-\delta_{M,\gamma}/2)} + \frac{\hat{\delta}_{M,\gamma}/2}{1-\gamma_{Bw}(1-\hat{\delta}_{M,\gamma})}\right). \quad (4)$$

Fig. 1 supports the idea that Thm. 1 becomes tighter than prior results when the variance due to model approximation (Def. 5) is low or when $\frac{R_{max}}{1-\gamma}$ is dominant (γ close to 1).

3 Bias under partial observability

We now look at how partial observability impacts the structural parameters to better understand its impact on the bias. This is important since most practical problems suffer from a form of partial observability (Dulac-Arnold et al., 2021). We consider a discrete-time POMDP (Sondik, 1978) described by the MDP tuple extended with two elements: a finite set of possible observations Ω and the probabilities of receiving each observation given a state, $\mathbb{O} : \mathcal{S} \times \Omega \mapsto [0, 1]$. On each time step $t \in \mathbb{N}_0$, the current state $S_t \in \mathcal{S}$ leads the agent to receive an observation $O_t \in \Omega$ (using $\mathbb{O}$), an action $A_t \in \mathcal{A}$ is played, the environment transitions into the next (unknown) state S_{t+1} (using P) and generates an observed reward R_{t+1} (using R).

When facing a partially observable setting, an effective way of approximating a solution is to use a policy defined on compressed histories (Francois-Lavet et al., 2019). Let $\mathcal{H}_t = \Omega \times (\mathcal{A} \times R \times \Omega)^t$ denote the set of histories observed up to time t and let $\mathcal{H} = \bigcup_{t=0}^{\infty} \mathcal{H}_t$ denote the space of all possible histories. The belief state $b(s|H)$ is a vector where the i-th component ($i \in \{1, \ldots, |\mathcal{S}|\}$) corresponds to $\mathbb{P}(s = i|H)$, for any history $H \in \mathcal{H}$. One can define a mapping $\phi : \mathcal{H} \mapsto \phi(\mathcal{H})$, where $\phi(\mathcal{H}) = \{\phi(H)|H \in \mathcal{H}\}$ is finite, which can be used as input to a policy $\pi : \phi(\mathcal{H}) \mapsto \mathcal{A}$.

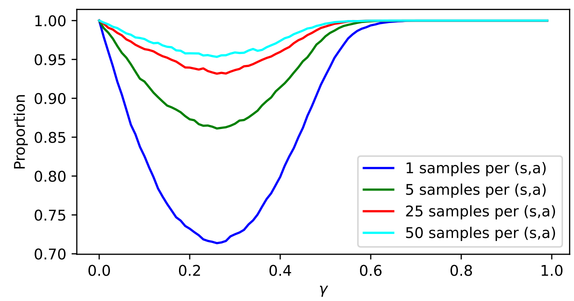

Figure 1: Proportion of randomly sampled MDPs where Eq. 4 is true given a discount factor γ.

Given a POMDP (extended tuple) M and any given distribution $\mathcal{D}_H$ over histories, one can define the expected return obtained over an infinite time horizon from a given history H, with $A_t \sim \pi(\phi(H_t))$:

$$V^{\pi}_{M,\gamma}(\phi(H)) = \mathop{\mathbb{E}}_{\substack{H' \sim \mathcal{D}_H : \\ \phi(H') = \phi(H)}} [V^{\pi}_{M}(H'|\phi)]$$

$$V^{\pi}_{M,\gamma}(H'|\phi) = \mathbb{E}_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t \sim b(\cdot|H_t = H'), \right].$$

For a given mapping ϕ, the optimal policy $\pi^*_{M,\gamma}$ maximizes $V^{\pi}_{M,\gamma}(\phi(H))$ for all histories $H \in \mathcal{H}$.

3.1 Extending structural parameters

We can extend Definitions 1 and 2 to the POMDP setting by applying them to compressed histories rather than the actual states in the underlying MDP:

$$\kappa^{\phi}_{M,\gamma} = \max_{\sigma,\sigma' \in \phi(\mathcal{H})} \left| V^{\pi^*_{M,\gamma}}_{M,\gamma}(\sigma) - V^{\pi^*_{M,\gamma}}_{M,\gamma}(\sigma') \right| \tag{5}$$

$$\delta^{\phi}_{M} = \max_{\sigma \in \phi(\mathcal{H})} \max_{a,a' \in \mathcal{A}} \sum_{\sigma' \in \phi(\mathcal{H})} |\mathbb{P}(\sigma'|\sigma, a) - \mathbb{P}(\sigma'|\sigma, a')| . \tag{6}$$

We introduce the following result showing how the structural parameters in the POMDP relate to the structural parameters of the underlying MDP (see Appendix F):

Theorem 2. *Given a POMDP M, let $\kappa^{\mathcal{S}}_{M,\gamma}$ and $\delta^{\mathcal{S}}_{M}$ denote the structural parameters (Definitions 1 and 2) evaluated on the underlying state space. Let $\mathcal{H}(s) = \bigcup_{t=0}^{\infty} \{H_t : b(s|H_t) > 0, H_t \in \mathcal{H}_t\}$ denote the set of all histories which can lead to being in state s at any time t, and $\mathcal{D}_H(s)$ denote its probability distribution. Define Δ^{ϕ}_{ϵ} s.t. $\left| V^{\pi^*_{M,\gamma}}_{M,\gamma}(s) - V^{\pi_{\phi}}_{M,\gamma}(s) \right| \leq \Delta^{\phi}_{\epsilon}\ \forall s \in \mathcal{S}$, define $\pi_{\phi}(s) = \mathop{\mathbb{E}}_{H \sim \mathcal{D}_H(s)} \pi^*_{M,\gamma}(\phi(H))$ as the optimal policy on compression of histories when executed over the underlying state space and define $b(s|\sigma) = \mathop{\mathbb{E}}_{H \sim \mathcal{D}_H(s), \sigma = \phi(H)} b(s|H)$ for $\sigma \in \phi(\mathcal{H})$. We have that:*

$$\delta^{\phi}_{M} \leq \delta^{\mathcal{S}}_{M} \qquad \textit{and} \qquad \kappa^{\phi}_{M,\gamma} \leq \max_{\sigma,\sigma' \in \phi(\mathcal{H})} \frac{\|b(\cdot|\sigma) - b(\cdot|\sigma')\|_1}{2} \left(\kappa^{\mathcal{S}}_{M,\gamma} + \Delta^{\phi}_{\epsilon} \right).$$

The first inequality confirms that partial observability impacts negatively the ability for the agent to control state transitions. The second inequality implies that if the policy on the partially observable domain remains good ($\Delta^{\phi}_{\epsilon} \approx 0$), then the state-value variation observed by the agent is lower (since the L_1 distance is bounded by 2), which could make the learning task easier and as efficient with a low discount factor. This L_1 distance often has a value of 2, which makes the bound quite loose, but they illustrate the idea that the values of structural parameters decrease when taking a convex combination over states. In fact the maximal value of the expectation of a random variable happens if all the mass is concentrated on the maximal value of the support. This is explained further in the appendix B. By considering that the horizon-sensitive action variation (Def. 4) is upper-bounded by the action variation (Def. 2), we can observe from Eq. 3 that the bias in the underlying MDP upper-bounds the bias of the POMDP when the optimal policy under partial observability is accurate on the true state space. This extends the ideas from Abel et al. (2016) where abstractions can make a problem much easier to learn while retaining good performance and in our case, lower the bias.

4 Numerical experiments

We now conduct a series of experiments to highlight the relationships between the planning horizon, the partial observability, and the structural parameters of the (underlying) MDP. See code online.

Random MDPs We consider the simulated environment of Jiang et al. (2016). We use 2-actions MDPs, with $Fixed(|\mathcal{S}|, d)$ denoting a randomly generated MDP with $d \geq 1$ next states reachable from each state. MDPs are sampled using the following procedure: 1) each state-action pair is assigned d possible next states; 2) transition probabilities to these states are sampled uniformly in $[0, 1]$, then normalized; 3) rewards are assigned to state-action pairs by sampling uniformly in $[0, 1]$.

Extension to partial observability We consider the state-abstraction setting (Abel et al., 2016), which corresponds to a specific case of partial observability where the history compressor $\phi(\mathcal{H})$ returns only the last observation and where $\mathbb{O}$ is a one-hot vector on an observation from Ω. For simplicity, we make sure that each observation is connected to at least one state. Using Bayes' theorem to recover the belief that the agent is in state s given observation ω, we get a constant uniform distribution on every state s which maps onto this observation:

$$b(s|\omega) = \frac{1}{|\{s \in \mathcal{S} : \mathbb{O}(o, s) > 0\}|} \forall s \in \mathcal{S} : \mathbb{O}(\omega, s) > 0, \forall \omega \in \Omega,$$

and a belief of 0 otherwise. From this special case of POMDP, we can extract an abstract MDP $M_A = \langle \mathcal{S}_A, \mathcal{A}, P_A, R_A, \gamma \rangle$ from the underlying MDP $M = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$ by using (Abel et al., 2016):

$$R_A(\omega, a) = \sum_{s \in \mathcal{S}} b(s|\omega) R(s, a) \tag{7}$$

$$P_A(\omega, a, \omega') = \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} P(s, a, s') b(s|\omega) \mathbb{O}(\omega', s'). \tag{8}$$

For our experiments under partial observability, we start by sampling an MDP from $Fixed(|\mathcal{S}|, d)$. Then, we can map it onto abstracted MDPs (POMDPs) with different number of observations is $|\Omega|$. The number of obervations encodes the level of partial observability. For $|\Omega| = |\mathcal{S}|$, the problem is fully observable. For $|\Omega| = 1$ (and $|\mathcal{S}| > 1$), the agent is completely blind to the state. We sample 10^4 MDPs with $Fixed(10, 3)$ and abstract each MDP into 6 configurations: $|\Omega| \in \{10, 8, 6, 4, 2, 1\}$. The Blackwell discount factors are computed by iterating from $\gamma = 1$ to $\gamma = 0$ with step size of 0.01 until the optimal policy changes.

Fig. 2 (left) shows that the mass of the Blackwell planning horizon tends to decrease as the observability decreases. Since the bias is null when planning with a discount factor larger than γ_{Bw}, we only cumulate variance above that point. When $|\Omega| = 1$, the myopic agent ($\gamma = 0$) enjoys the optimal planning horizon, which corresponds to the bandit setting.

We also evaluate the normalized bias $\max_{s \in \mathcal{S}} \left(V_{M,\gamma_{Bw}}^{\pi^*_{M,\gamma_{Bw}}}(s) - V_{M,\gamma_{Bw}}^{\pi^*_{M,\gamma}}(s) \right) / V_{M,\gamma_{Bw}}^{\pi^*_{M,\gamma_{Bw}}}(s)$ for different planning horizons, averaged over different levels of partial observability. Fig. 2 (right) shows that although the bias decreases when the planning horizon increases, this effect attenuates as the observability decreases. Given that many real-world problems are partially observable, this finding supports the need to consider shallow planning more seriously.

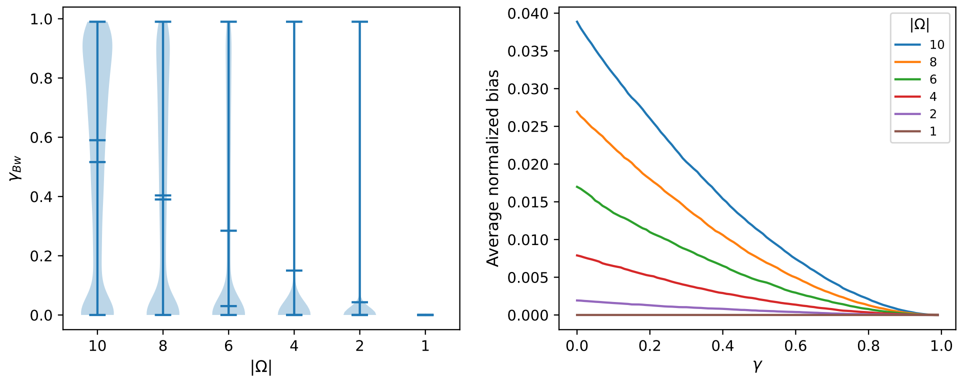

Figure 2: Left: Distribution of Blackwell discount factors over 10^4 POMDPs given the number of observations. Right: Average normalized bias given the planning horizon and number of observations.

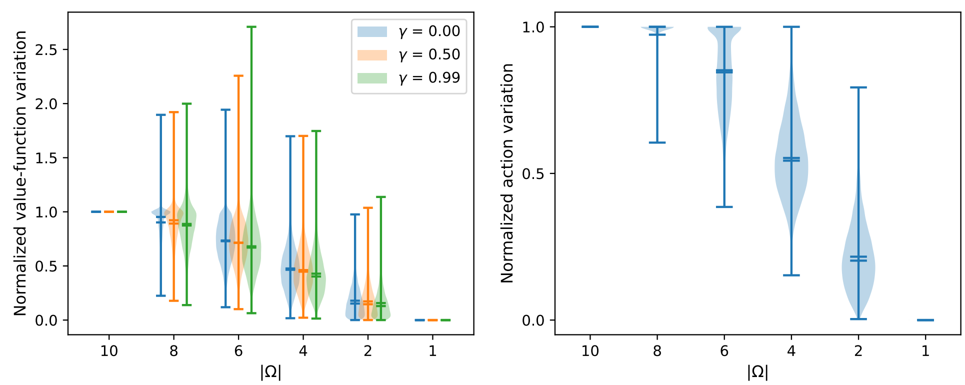

Figure 3: Left: Distribution of normalized $\kappa^{\phi}_{M,\gamma}$ over 10^4 POMDPs given the number of observations and the discount factor used. Right: Distribution of normalized δ^{ϕ}_{M} given the number of observations.

We normalize the parameters by dividing the structural parameters for each abstract MDP by the underlying parameter in the true $Fixed(\cdot)$ MDP. Fig. 3 (right) highlights the strict inequality offered by Thm. 2 on the action variation of the POMDP vs the underlying MDP. By observing Thm. 1 and Fig. 3, the reduction in structural parameters offers insight into how why the bias decreases under partial observability. It also points to the fact that there might exist a bound on the Blackwell discount factor using structural parameters much tighter than the one provided in Grand-Clément and Petrik (2023) or even improve upon the results in Laidlaw et al. (2023) using partial observability.

5 Conclusion

We extended existing structural parameters to consider the planning horizon (Def. 4) and the model approximation (Def. 6). This allowed us to extend an existing bound on the bias (Eq. 3) and propose a new bound on the variance (Lemma 1), which resulted in a new bound on the planning loss (Thm. 1). We finally extended the structural parameters to POMDPs (Eq. 5 and 6) and showed that these are controlled by their fully observable counterparts (Thm. 2). This complements previous results (Abel et al., 2016; Francois-Lavet et al., 2019) by considering the impact of the planning horizon on the bias when shallow planning under partial observability.

References

Abel, D., Hershkowitz, D., and Littman, M. (2016). Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*.

Amit, R., Meir, R., and Ciosek, K. (2020). Discount Factor as a Regularizer in Reinforcement Learning. In *International Conference on Machine Learning*.

Bastani, H. and Bayati, M. (2020). Online decision making with high-dimensional covariates. *Operations Research*.

Devidze, R., Radanovic, G., Kamalaruban, P., and Singla, A. (2021). Explicable reward design for reinforcement learning agents. *Advances in Neural Information Processing Systems (NeurIPS)*.

Ding, K., Li, J., and Liu, H. (2019). Interactive anomaly detection on attributed networks. In *ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery.

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., and Hester, T. (2021). An empirical investigation of the challenges of real-world reinforcement learning. *Machine Learning*.

Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D., and Pineau, J. (2018). Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine Learning for Healthcare Conference*.

Francois-Lavet, V., Rabusseau, G., Pineau, J., Ernst, D., and Fonteneau, R. (2019). On Overfitting and Asymptotic Bias in Batch Reinforcement Learning with Partial Observability. *Journal of Artificial Intelligence Research (JAIR)*.

Gheshlaghi Azar, M., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*.

Grand-Clément, J. and Petrik, M. (2023). Reducing blackwell and average optimality to discounted MDPs via the blackwell discount factor. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., Yang, Y., and Knoll, A. (2023). A review of safe reinforcement learning: Methods, theory and applications. *arXiv:2205.10330*.

Hare, J. (2019). Dealing with sparse rewards in reinforcement learning. *arXiv:1910.09281*.

He, J., Zhou, D., and Gu, Q. (2021). Uniform-PAC bounds for reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems (NeurIPS)*.

Jiang, N., Kulesza, A., Singh, S., and Lewis, R. (2015). The dependence of effective planning horizon on model accuracy. In *International Conference on Autonomous Agents and Multiagent Systems*.

Jiang, N., Singh, S., and Tewari, A. (2016). On structural properties of MDPs that bound loss due to shallow planning. In *International Joint Conference on Artificial Intelligence*.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*.

Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R. H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., Mohiuddin, A., Sepassi, R., Tucker, G., and Michalewski, H. (2024). Model-based reinforcement learning for atari. *arXiv:1903.00374*.

Laidlaw, C., Russell, S. J., and Dragan, A. (2023). Bridging RL theory and practice with the effective horizon. *Advances in Neural Information Processing Systems (NeurIPS)*.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv:1312.5602*.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 1 edition.

Sondik, E. J. (1978). The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations research*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition edition.

Wu, Y., He, J., and Gu, Q. (2023). Uniform-PAC guarantees for model-based rl with bounded eluder dimension. In *Uncertainty in Artificial Intelligence*.

Yu, Y. (2018). Towards sample efficient reinforcement learning. In *International Joint Conference on Artificial Intelligence*.

A Notation

Symbol	**Description**
$\mathcal{S}$	Finite state space
$\mathcal{A}$	Finite action space
P	Transition function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$
R	Reward function $R : \mathcal{S} \times \mathcal{A} \mapsto [0, R_{\max}]$
γ_{Bw}	Blackwell discount factor
$V^{\pi}_{M,\gamma}$	Vector of state values when following policy π on MDP M with discount factor γ
$\pi^{*}_{M,\gamma}$	Optimal policy in MDP M with discount factor γ
$\mathcal{Z}_M(\pi \neq \pi')$	Set of state-action pairs on MDP M where two policies π and π' differ
$\hat{\epsilon}$	Variance due to model approximation
δ_M	Action variation for MDP M
$\kappa_{M,\gamma}$	Value-function variation in MDP M with discount factor γ
$\delta_{M,\gamma}$	Horizon-sensitive action variation in MDP M with discount factor γ
$\hat{\delta}_{M,\gamma}$	Empirical action variation in MDP M with discount factor γ
$P^{\pi}_{s,k}$	$\lvert\mathcal{S}\rvert \times \lvert\mathcal{S}\rvert$ matrix of transition probabilities from state s when following policy π for k steps
$[P^{\pi}_{M}]$	$\lvert\mathcal{S}\rvert \times \lvert\mathcal{S}\rvert$ matrix of transition probabilities when following policy π in MDP M
$e^{\top}_{s}$	One-hot vector of length $\lvert\mathcal{S}\rvert$ with a 1 at the index of $s \in \mathcal{S}$
Ω	Finite set of observations
$\mathbb{O}$	Observation probabilities $\mathbb{O} : \mathcal{S} \times \Omega \mapsto [0, 1]$
$\mathcal{H}_t$	Set of histories observed up to time t
$\mathcal{H}$	Space of all possible histories
$\mathcal{D}_H$	Distribution over histories
$b(s \vert H)$	Belief state, $\mathbb{P}(s = i \vert H)$ for history $H \in \mathcal{H}$
ϕ	Mapping from histories to a finite set, $\phi : \mathcal{H} \mapsto \phi(\mathcal{H})$
$\phi(\mathcal{H})$	Space of all possible compression of histories
$\pi(\phi(H))$	Policy defined on compressed histories
$\pi_{\phi}(s)$	Optimal policy defined over compressed histories executed on the true state space
$V^{\pi}_{M,\gamma}(\phi(H))$	Expected return from history H with policy π
$\pi^{*}_{M,\gamma}$	Optimal policy for POMDP M and discount factor γ
Δ^{ϕ}_{ϵ}	Loss of performance on the true state space from executing $\pi_{\phi}(s)$ instead of the optimal policy
$\kappa^{\phi}_{M,\gamma}$	Value-function variation in POMDP M with discount factor γ and mapping ϕ
δ^{ϕ}_{M}	Action variation in POMDP M and mapping ϕ

Table 1: List of notations

B Proof of Eq. 3

Let $[P^{\pi}_{M}]$ denote a $|\mathcal{S}| \times |\mathcal{S}|$ matrix with the element indexed by (s, s') being $P(s'|s,a)$ under policy π and $e_s^\top$ be a one-hot vector on an arbitrary state $s \in \mathcal{S}$, that is a vector of 0s and a value of 1 for the index s. We use the following result from Appendix A, Sec. 1 in Jiang et al. (2016):

Fact 1 (State value decomposition, Jiang et al. (2016)). *Let M be and MDP with Blackwell discount factor $\gamma_{Bw} > \gamma$:*

$$V^{\pi}_{M,\gamma_{Bw}}(s) = e_s^\top V^{\pi}_{M,\gamma} + \sum_{k=1}^{\infty} (\gamma_{Bw} - \gamma)\gamma_{Bw}^{k-1} e_s^\top [P^{\pi}_{M}]^k V^{\pi}_{M,\gamma}.$$

We then have for any arbitrary state $s \in \mathcal{S}$ by using fact 1 on both state values of the bias:

$$\begin{aligned}
V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}}(s) - V^{\pi^*_{M,\gamma}}_{M,\gamma_{Bw}}(s) &= e_s^\top V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma} - e_s^\top V^{\pi^*_{M,\gamma}}_{M,\gamma} \\
&\quad + (\gamma_{Bw} - \gamma) \sum_{k=1}^{\infty} \gamma_{Bw}^{k-1} \left(e_s^\top [P^{\pi_{M,\gamma_{Bw}}}_{M}]^k V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma} - e_s^\top [P^{\pi_{M,\gamma}}_{M}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} \right). \\
&\leq 0 + (\gamma_{Bw} - \gamma) \sum_{k=1}^{\infty} \gamma_{Bw}^{k-1} \left(e_s^\top [P^{\pi_{M,\gamma_{Bw}}}_{M}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} - e_s^\top [P^{\pi_{M,\gamma}}_{M}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} \right).
\end{aligned}$$

By realizing that $V^{\pi^*_{M,\gamma}}_{M,\gamma}(s) \geq V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma}(s)\ \forall s \in \mathcal{S}$, the first term is upper bounded by 0. We can also start by bounding the inner term of the sum. To achieve this, we will use the following result:

Lemma 2 (from Jiang et al. (2016)). *Given stochastic vectors $p, q \in \mathbb{R}^{|S|}$, and a real vector V with the same dimension,*

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

In a way, this lemma gives a bound on the difference between two expectations defined over the same support. The result is quite loose since it doesn't take well into consideration the form of the distribution in p and q. We will still use it since its the only result we know, and will still be useful to illustrate our ideas.

We can bound the inner sum using Lemma 2:

$$e_s^\top [P^{\pi_{M,\gamma_{Bw}}}_{M}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} - e_s^\top [P^{\pi_{M,\gamma}}_{M}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} \leq \left\| e_s^\top [P^{\pi_{M,\gamma_{Bw}}}_{M}]^k - e_s^\top [P^{\pi_{M,\gamma}}_{M}]^k \right\|_1 \frac{\kappa_{M,\gamma}}{2}.$$

In their proof Jiang et al. (2016) use the action variation as an upper bound to the row-wise L_1-norm distance between the transition matrices $[P^{\pi_1}_{M}]$ and $[P^{\pi_2}_{M}]$ of two policies π_1 and π_2. By realizing that this distance is 0 on states where $\pi_1(s) = \pi_2(s)$, we can tighten up the bounds using Def. 3, changing a single step in the proof of Lemma 1 from Jiang et al. (2016), and using their Corollary 4 with our new maximal l_1-norm distance in definition 4 and definition 6 to get the following inequalities, which are equivalent to Propositions 1 and 2 respectively:

$$\left\| e_s^\top [P^{\pi^*_{M,\gamma}}_{M}]^k - e_s^\top [P^{\pi^*_{M,\gamma_{Bw}}}_{M}]^k \right\|_1 \leq 2 - 2(1 - \delta_{M,\gamma}/2)^k$$

$$\left\| e_s^\top [P^{\pi^*_{M,\gamma}}_{M}]^k - e_s^\top [P^{\pi^*_{\hat{M},\gamma}}_{M}]^k \right\|_1 \leq 2 - 2(1 - \hat{\delta}_{M,\gamma}/2)^k.$$

By plugging the second inequality in our inner term, we obtain the following bound:

$$\begin{aligned}
V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}}(s) - V^{\pi^*_{M,\gamma}}_{M,\gamma_{Bw}}(s) &\leq \sum_{k=1}^{\infty} (\gamma_{Bw} - \gamma){\gamma_{Bw}}^{k-1} \big(1 - (1 - \delta_{M,\gamma}/2)^k\big) \kappa_{M,\gamma} \\
&= \frac{\delta_{M,\gamma}/2 \cdot \kappa_{M,\gamma}(\gamma_{Bw} - \gamma)}{(1 - \gamma_{Bw})(1 - \gamma_{Bw}(1 - \delta_{M,\gamma}/2))}.
\end{aligned}$$

□

C Proof of Lemma 1

For this proof, we will use the same setting as section B. Using Fact 1 on both state values in the variance, we have the following:

$$\underbrace{V^{\pi^*_{M,\gamma}}_{M,\gamma_{Bw}}(s) - V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma_{Bw}}(s)}_{\text{variance}} = e_s^\top V^{\pi^*_{M,\gamma}}_{M,\gamma} - e_s^\top V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma} + (\gamma_{Bw} - \gamma) \sum_{k=1}^{\infty} \gamma_{Bw}^{k-1} \left(e_s^\top [P_M^{\pi_{M,\gamma}}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} - e_s^\top [P_M^{\pi_{\hat{M},\gamma}}]^k V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma} \right). \tag{9}$$

We will start by bounding the sum first. The inner term can be rewritten as:

$$e_s^\top [P_M^{\pi_{M,\gamma}}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} - e_s^\top [P_M^{\pi_{\hat{M},\gamma}}]^k V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma} = e_s^\top [P_M^{\pi_{M,\gamma}}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} - e_s^\top [P_M^{\pi_{\hat{M},\gamma}}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} + e_s^\top [P_M^{\pi_{\hat{M},\gamma}}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} - e_s^\top [P_M^{\pi_{\hat{M},\gamma}}]^k V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma}. \tag{10}$$

Note that there is a loss in tightness that comes from the fact that to isolate $\hat{\epsilon}$ (to use the existing litterature on the variance term), we need to use this decomposition, on which we will then use Holder's inequality. For the next results, we use Lemma 2 on the first difference, and using Holder's inequality on the second, we obtain:

$$\begin{aligned} e_s^\top [P_M^{\pi_{M,\gamma}}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} - e_s^\top [P_M^{\pi_{\hat{M},\gamma}}]^k V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma} &\leq \left\| e_s^\top [P_M^{\pi_{M,\gamma}}]^k - e_s^\top [P_M^{\pi_{\hat{M},\gamma}}]^k \right\|_1 \frac{\kappa_{M,\gamma}}{2} + \left\| V^{\pi^*_{M,\gamma}}_{M,\gamma} - V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma} \right\|_\infty \\ &= \left\| e_s^\top [P_M^{\pi_{M,\gamma}}]^k - e_s^\top [P_M^{\pi_{\hat{M},\gamma}}]^k \right\|_1 \frac{\kappa_{M,\gamma}}{2} + \hat{\epsilon}. \end{aligned} \tag{11}$$

For bounding the first term, we use the inequality presented in section B.Then, we have that the term inside the sum from Eq 9 is bounded by:

$$e_s^\top [P_M^{\pi_{M,\gamma}}]^k V^{\pi^*_{M,\gamma}}_{M,\gamma} - e_s^\top [P_M^{\pi_{\hat{M},\gamma}}]^k V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma} \leq \frac{(2 - 2(1 - \hat{\delta}_{M,\gamma}/2)^k)\kappa_{M,\gamma}}{2} + \hat{\epsilon}.$$

This allows us to obtain the following bound on the variance:

$$\begin{aligned} \underbrace{V^{\pi^*_{M,\gamma}}_{M,\gamma_{Bw}}(s) - V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma_{Bw}}(s)}_{\text{variance}} &\leq \hat{\epsilon} + \sum_{k=1}^{\infty} (\gamma_{Bw} - \gamma) {\gamma_{Bw}}^{k-1} \left(1 - (1 - \hat{\delta}_{M,\gamma}/2)^k\right) \kappa_{M,\gamma} + \frac{(\gamma_{Bw} - \gamma)\hat{\epsilon}}{1 - \gamma_{Bw}} \\ &= \hat{\epsilon} \left(\frac{1 - \gamma}{1 - \gamma_{Bw}} \right) + \frac{\hat{\delta}_{M,\gamma}/2 \cdot \kappa_{M,\gamma}(\gamma_{Bw} - \gamma)}{(1 - \gamma_{Bw})(1 - \gamma_{Bw}(1 - \hat{\delta}_{M,\gamma}/2))}. \end{aligned}$$

□

D Proof of Theorem 1

Recall the following decomposition of the planning loss (Eq. 1):

$$||V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}} - V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma_{Bw}}||_\infty \leq \underbrace{||V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}} - V^{\pi^*_{M,\gamma}}_{M,\gamma_{Bw}}||_\infty}_{\text{bias}} + \underbrace{||V^{\pi^*_{M,\gamma}}_{M,\gamma_{\mathrm{Bw}}} - V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma_{Bw}}||_\infty}_{\text{variance}}.$$

We can upper-bound the bias using a result from Jiang et al. (2016), which we extended (see Appendix 2.1) to explicit the dependence on the planning horizon using Def. 4:

$$||\underbrace{V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}} - V^{\pi^*_{M,\gamma}}_{M,\gamma_{Bw}}}_{\text{bias}}||_\infty \leq \frac{\delta_{M,\gamma}/2 \cdot \kappa_{M,\gamma}(\gamma_{Bw} - \gamma)}{(1 - \gamma_{Bw})(1 - \gamma_{Bw}(1 - \delta_{M,\gamma}/2))}.$$

We can combine the bias upper bound with Lemma 1 and obtain Thm. 1:

$$
\begin{aligned}
\|V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}} - V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma_{Bw}}\|_\infty &\leq \underbrace{\|V^{\pi^*_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}} - V^{\pi^*_{M,\gamma}}_{M,\gamma_{Bw}}\|_\infty}_{\text{bias}} + \underbrace{\|V^{\pi^*_{M,\gamma}}_{M,\gamma_{\text{Bw}}} - V^{\pi^*_{\hat{M},\gamma}}_{M,\gamma_{Bw}}\|_\infty}_{\text{variance}}. \\
&\leq \frac{\delta_{M,\gamma}/2 \cdot \kappa_{M,\gamma}(\gamma_{Bw} - \gamma)}{(1-\gamma_{Bw})(1-\gamma_{Bw}(1-\delta_{M,\gamma}/2))} + \hat{\epsilon}\left(\frac{1-\gamma}{1-\gamma_{Bw}}\right) + \frac{\hat{\delta}_{M,\gamma}/2 \cdot \kappa_{M,\gamma}(\gamma_{Bw} - \gamma)}{(1-\gamma_{Bw})(1-\gamma_{Bw}(1-\hat{\delta}_{M,\gamma}/2))} \\
&= \kappa_{M,\gamma}\left(\frac{\gamma_{Bw} - \gamma}{1-\gamma_{Bw}}\right)\left(\frac{\delta_{M,\gamma}/2}{1-\gamma_{Bw}(1-\delta_{M,\gamma}/2)} + \frac{\hat{\delta}_{M,\gamma}/2}{1-\gamma_{Bw}(1-\hat{\delta}_{M,\gamma}/2)}\right) + \hat{\epsilon}\left(\frac{1-\gamma}{1-\gamma_{Bw}}\right),
\end{aligned}
$$

where the last equality is obtained by rearranging the terms.

E Tightness of Theorem 1

The bound offered by Thm. 1 is tighter than prior results (Jiang et al., 2015) when the gain in tightness from using structural parameters is higher than the loss incurred from having a looser variance term in Eq. 1. Formally, we are looking for the condition such that Thm. 1 is tighter than prior results (Jiang et al., 2015):

$$
\kappa_{M,\gamma}\left(\frac{\gamma_{Bw} - \gamma}{1-\gamma_{Bw}}\right)\left(\frac{\delta_{M,\gamma}/2}{1-\gamma_{Bw}(1-\delta_{M,\gamma}/2)} + \frac{\hat{\delta}_{M,\gamma}/2}{1-\gamma_{Bw}(1-\hat{\delta}_{M,\gamma}/2)}\right) + \hat{\epsilon}\left(\frac{1-\gamma}{1-\gamma_{Bw}}\right) \leq \frac{\gamma_{Bw} - \gamma}{(1-\gamma_{Bw})(1-\gamma)} R_{\max} + \hat{\epsilon},
$$

which we can re-arrange until we obtain Eq. 4:

$$
\begin{aligned}
&\kappa_{M,\gamma}\left(\frac{\delta_{M,\gamma}/2}{1-\gamma_{Bw}(1-\delta_{M,\gamma}/2)} + \frac{\hat{\delta}_{M,\gamma}/2}{1-\gamma_{Bw}(1-\hat{\delta}_{M,\gamma}/2)}\right) + \hat{\epsilon}\left(\frac{1-\gamma}{\gamma_{Bw}-\gamma}\right) \leq \frac{R_{\max}}{(1-\gamma)} + \hat{\epsilon}\left(\frac{1-\gamma_{Bw}}{\gamma_{Bw}-\gamma}\right) \\
&\kappa_{M,\gamma}\left(\frac{\delta_{M,\gamma}/2}{1-\gamma_{Bw}(1-\delta_{M,\gamma}/2)} + \frac{\hat{\delta}_{M,\gamma}/2}{1-\gamma_{Bw}(1-\hat{\delta}_{M,\gamma}/2)}\right) - \frac{R_{\max}}{(1-\gamma)} \leq \hat{\epsilon}\left(\frac{1-\gamma_{Bw}}{\gamma_{Bw}-\gamma}\right) - \hat{\epsilon}\left(\frac{1-\gamma}{\gamma_{Bw}-\gamma}\right) \\
&\hat{\epsilon} \leq \frac{R_{max}}{1-\gamma} - \kappa_{M,\gamma}\left(\frac{\delta_{M,\gamma}/2}{1-\gamma_{Bw}(1-\delta_{M,\gamma}/2)} + \frac{\hat{\delta}_{M,\gamma}/2}{1-\gamma_{Bw}(1-\hat{\delta}_{M,\gamma})}\right).
\end{aligned}
$$

□

F Proof of Theorem 2

Let us begin with the result on the action variation in Thm. 2. We can decompose Eq. 6 and find the underlying structural parameter on the true state space. We start this by realizing that $\mathbb{P}(\sigma'|\sigma, a)$ is an expected value over possible histories distributed under an arbitrary distribution over histories $\mathcal{D}_H$. Then we decompose the transition probability into its underlying components using the belief state, as done in Francois-Lavet et al. (2019):

$$
\begin{aligned}
\delta_M^\phi &= \max_{\sigma \in \phi(\mathcal{H})} \max_{a,a' \in \mathcal{A}} \sum_{\sigma' \in \phi(\mathcal{H})} |\mathbb{P}(\sigma'|\sigma, a) - \mathbb{P}(\sigma'|\sigma, a')| \\
&= \max_{\sigma \in \phi(\mathcal{H})} \max_{a,a' \in \mathcal{A}} \sum_{\sigma' \in \phi(\mathcal{H})} \left| \underset{\substack{H' \sim \mathcal{D}_H: \\ \phi(H')=\sigma}}{E} \mathbb{P}(\sigma'|H', a) - \underset{\substack{H' \sim \mathcal{D}_H: \\ \phi(H')=\sigma}}{E} \mathbb{P}(\sigma'|H', a') \right| \\
&= \max_{\sigma \in \phi(\mathcal{H})} \max_{a,a' \in \mathcal{A}} \sum_{\sigma' \in \phi(\mathcal{H})} \left| \underset{\substack{H' \sim \mathcal{D}_H: \\ \phi(H')=\sigma}}{E} \mathbb{P}(\sigma'|H', a) - \mathbb{P}(\sigma'|H', a') \right| \\
&= \max_{\sigma \in \phi(\mathcal{H})} \max_{a,a' \in \mathcal{A}} \sum_{\sigma' \in \phi(\mathcal{H})} \left| \underset{\substack{H' \sim \mathcal{D}_H: \\ \phi(H')=\sigma}}{E} \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} b(s|H') p(s'|s, a) p(\sigma'|s', H') \right. \\
&\quad \left. - \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} b(s|H') p(s'|s, a') p(\sigma'|s', H') \right| \\
&= \max_{\sigma \in \phi(\mathcal{H})} \max_{a,a' \in \mathcal{A}} \sum_{\sigma' \in \phi(\mathcal{H})} \left| \underset{\substack{H' \sim \mathcal{D}_H: \\ \phi(H')=\sigma}}{E} \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \left(p(s'|s, a) - p(s'|s, a')\right) b(s|H') p(\sigma'|s', H') \right| \\
&\leq \max_{\sigma \in \phi(\mathcal{H})} \max_{a,a' \in \mathcal{A}} \sum_{\sigma' \in \phi(\mathcal{H})} \underset{\substack{H' \sim \mathcal{D}_H: \\ \phi(H')=\sigma}}{E} \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} |p(s'|s, a) - p(s'|s, a')| \, b(s|H') p(\sigma'|s', H') \\
&\leq \delta_M^S \max_{\sigma \in \phi(\mathcal{H})} \sum_{\sigma' \in \phi(\mathcal{H})} \underset{\substack{H' \sim \mathcal{D}_H: \\ \phi(H')=\sigma}}{E} p(\sigma'|s', H') \\
&= \delta_M^S.
\end{aligned}
$$

We obtain the first inequality by using the triangle inequality and on the second, we use Holder's inequality on the dot product for $s \in \mathcal{S}$ to retrieve δ_M^S. On this last inequality, we also interchange the order of summations, summing all probabilities on the support of σ' and then taking the expectation of the constant 1, which gives the result.

We take a similar approach to obtain the result on the value-function variation in Thm. 2. We can decompose Eq. 5 to retrieve the equivalent parameter on the true state space:

$$
\begin{aligned}
\kappa_{M,\gamma}^\phi &= \max_{\sigma, \sigma' \in \phi(\mathcal{H})} \left| V_{M,\gamma}^{\pi^*_{M,\gamma}}(\sigma) - V_{M,\gamma}^{\pi^*_{M,\gamma}}(\sigma') \right| \\
&= \max_{\sigma, \sigma' \in \phi(\mathcal{H})} \left| \sum_{s \in \mathcal{S}} b(s|\sigma) V_{M,\gamma}^{\pi_\phi}(s) - \sum_{s \in \mathcal{S}} b(s|\sigma') V_{M,\gamma}^{\pi_\phi}(s) \right| \\
&\leq \max_{\sigma, \sigma' \in \phi(\mathcal{H})} \frac{\|b(\cdot|\sigma) - b(\cdot|\sigma')\|_1}{2} \max_{s, s' \in \mathcal{S}} \left| V_{M,\gamma}^{\pi_\phi}(s) - V_{M,\gamma}^{\pi_\phi}(s') \right| \\
&\leq \max_{\sigma, \sigma' \in \phi(\mathcal{H})} \frac{\|b(\cdot|\sigma) - b(\cdot|\sigma')\|_1}{2} \left(\kappa_{M,\gamma}^{\mathcal{S}} + \Delta_\epsilon^\phi\right)
\end{aligned}
$$

We obtain the first inequality by using lemma 2. The last one is obtained by observing that $V_{M,\gamma}^{\pi^*_{M,\gamma}}(s) \geq V_{M,\gamma}^{\pi_\phi}(s), \forall s \in \mathcal{S}$ and that by the assumption of Thm. 2 we have that $-V_{M,\gamma}^{\pi_\phi}(s) \leq \Delta_\epsilon^\phi - V_{M,\gamma}^{\pi^*_{M,\gamma}}(s), \forall s \in \mathcal{S}$. □