Optimal discounting for offline Input-Driven MDP

Randy Lefebvre, Audrey Durand

Keywords: Offline RL, Input-Driven MDP, Bias-variance tradeoff, Discount factor

Summary

Offline reinforcement learning has gained a lot of popularity for its potential to solve industry challenges. However, real-world environments are often highly stochastic and partially observable, leading long-term planners to overfit to offline data in model-based settings. Input-Driven Markov Decision Processes (IDMDPs) offer a way to work with some of the uncertainty by letting designers separate what the agent has control over (states) from what it cannot (inputs) in the environnement. These stochastic external inputs are often difficult to model. Under the assumption that the input model will be imperfect, we investigate the bias-variance tradeoff under shallow planning in IDMDPs. Paving the way to input-driven planning horizons, we also investigate the similarity of optimal planning horizons at different inputs given the structure of the input space.

Contribution(s)

1. We provide new insights connecting the input structure to the state-value function in Input-Driven MDPs (Lemma 1).
   Context: This result is also applicable to MDPs and therefore generalizes the value function variation from Jiang et al. (2016) to any policy and any pair of states.
2. We provide a novel bound on the variance due to the error in the input model and the planning horizon in offline Input-Driven MDPs (Lemma 2), which we use to obtain the first existing bound on the planning loss for Exo-MDPs (Theorem 1).
   Context: Prior results (Jiang et al., 2015; Lefebvre & Durand, 2025) study the variance due to the error in the state model in a MDP, i.e. considering variables that the agent can control (whereas the agent cannot control the inputs).
3. We provide the first results on the optimal input-dependent discount factor in Input-Driven MDPs. We connect the planning loss at different inputs to the input structure (Lemma 3), allowing to control the variation of optimal input-dependent discount factors over the input space using the input structure (Theorem 2).
   Context: This connects to the (limited) work on state-dependent discount factors, focusing on the impact of the non-controllable variables (inputs) on the optimal planning horizon.

Optimal discounting for offline Input-Driven MDP

Randy Lefebvre[1], **Audrey Durand**[1,2]
randy.lefebvre.1@ulaval.ca, audrey.durand@ift.ulaval.ca

[1]**Université Laval**
[2]**Canada CIFAR AI Chair**

Abstract

Offline reinforcement learning has gained a lot of popularity for its potential to solve industry challenges. However, real-world environments are often highly stochastic and partially observable, leading long-term planners to overfit to offline data in model-based settings. Input-Driven Markov Decision Processes (IDMDPs) offer a way to work with some of the uncertainty by letting designers separate what the agent has control over (states) from what it cannot (inputs) in the environnement. These stochastic external inputs are often difficult to model. Under the assumption that the input model will be imperfect, we investigate the bias-variance tradeoff under shallow planning in IDMDPs. Paving the way to input-driven planning horizons, we also investigate the similarity of optimal planning horizons at different inputs given the structure of the input space.

1 Introduction

Reinforcement learning (RL) has attracted significant attention for its ability to solve high-dimensional control problems, as demonstrated in simulators and video games (Mnih et al., 2015; Niu et al., 2022; Silver et al., 2016). Despite these successes, online RL remains difficult to deploy in industrial applications due to challenges like partial observability, security risks, and business constraints (Dulac-Arnold et al., 2021). Offline RL offers an alternative by learning policies from pre-collected data that reflects existing business operations and is easier to operationalize due to its similarity to conventional machine learning (Agarwal et al., 2020; Levine et al., 2020). To reduce risks, methods such as conservative Q-learning and expert-supervised RL constrain the learned policies to remain close to the training data (Kumar et al., 2020; Sonabend et al., 2020). For operational constraints (maximum budgets, time frames, etc.), another enticing setting for practical applications is model-based offline RL, where independent models can be developed to capture the different dynamics of the environment (Yu et al., 2020). These dynamics can then be adapted to match operational settings, leading to much more controllable and interpretable policies (Argenson & Dulac-Arnold, 2021). However, when applied to domains such as healthcare, finance, insurance, e-commerce, or social media, model-based offline RL still faces the significant challenge of partial observability: Critical state information is often missing, leading to stochastic observations with high aleatoric uncertainty resulting in poor policy generalization (Ghosh et al., 2021).

It has been shown that the bias-variance tradeoff can be improved by enabling the agent to perform shallow planning under partial observability or low data regimes (Jiang et al., 2015; Amit et al., 2020; Liu & Li, 2021; Cannelli et al., 2023; Lefebvre & Durand, 2025) In industry settings, some dimensions of the state-space are often highly stochastic and hard to model, but they get blended with other dimensions that are easier to model which makes it hard to know if shallow planning would help generalization. Therefore, in this work, we study the impact of shallow planning under the Input-Driven Markov Decision Process (IDMDP) setting (Mao et al., 2018) where state variables controlled by the agent (states) are modelled independently of those that the agent does not control

(inputs). The inputs are often rich observations which depend on external stochastic processes that are very hard to model accurately. One can then look at the error in inputs modelling to guide better choices of discount factor to improve generalization.

Contributions We introduce novel theoretical results on the bias-variance tradeoff induced by a shallow planning horizon in IDMDPs given an imperfect input dynamics model. By leveraging the structure of the input space, we then derive the first theoretical results on input-dependent optimal discount factors, motivating planning horizons adapted to local uncertainty in the input model. We support and illustrate our results using controlled numerical experiments and validate their generalization on a deep RL experiment. All our implementations are available for reproducibility[1].

2 Problem setting: Offline Input-Driven Markov Decision Processes

Markov Decision Process (MDP) We define a MDP by a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ 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 the transition function between states, and $R : \mathcal{S} \times \mathcal{A} \mapsto [0, R_{\max}]$ is the expected reward function. At each time step $t \in \mathbb{N}_0$ the current state $S_t \in \mathcal{S}$ is observed and the agent performs action $A_t \in \mathcal{A}$ according to it's policy $\pi : \mathcal{S} \mapsto \mathcal{A}$. Given this action, the environment transitions to the next state $\mathcal{S}_{t+1}$ using the transition function P and the agent receives the reward R_{t+1} using the reward function (given S_{t+1}). Given an MDP M, the value of state $s \in \mathcal{S}$ under policy π is the expected sum of discounted rewards by following actions under 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 planning horizon as $1/(1-\gamma)$ by assigning credit to future rewards in current state value. The goal of an RL agent is to find the optimal policy, i.e. the policy $\pi^{\star}_{M,\gamma} := \operatorname{argmax}_{\pi} V^{\pi}_{M,\gamma}(s)$ maximizing the value for all states $s \in \mathcal{S}$.

Input-Driven Markov Decision Process (IDMDP) IDMDPs were introduced to model environments where exogenous stochastic processes influence the underlying dynamics of an MDP (Mao et al., 2018). These external processes often transition in ways that differ significantly from the underlying states. For example, consider a hospital that manages bed allocation during a pandemic. The arrival of patients follows an external stochastic process driven by disease spread patterns, which the hospital cannot control but must account for in its decision-making. Similarly, a streaming service aiming to maximize long-term user engagement makes recommendations based on user characteristics and preferences, which evolve over time due to external influences such as social trends and personal life events, factors beyond the service's control. This distinction between information that the agent can and cannot control is fundamental to IDMDPs.

An IDMDP therefore extends the definition of an MDP by considering the arrival of *inputs* that are not controllable by the agent, but that can influence state transitions and the rewards. Formally, an IDMDP is defined by a tuple $(\mathcal{S}, \mathcal{Z}, \mathcal{A}, P_s, P_z, R, \gamma)$ where $\mathcal{S}$ and $\mathcal{A}$ respectively denote the finite state and action spaces as in a standard MDP, $\mathcal{Z}$ is a finite input space, $P_s : \mathcal{Z} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is the transition function between states, $P_z : \mathcal{Z} \times \mathcal{Z} \mapsto [0, 1]$ is the transition function between inputs, and $R : \mathcal{Z} \times \mathcal{S} \times \mathcal{A} \mapsto [0, R_{\max}]$ is the expected reward function. One observes that state transitions and expected rewards depend on both the current input, state, and action, while input transitions depend only on the current input. At each time step $t \in \mathbb{N}_0$, the agent observes the current input Z_t along with the current state S_t. The agent performs an action A_t given its policy $\pi : \mathcal{Z} \times \mathcal{S} \mapsto \mathcal{A}$. The system then transitions to the next state S_{t+1} using the transition function $P_s(S_{t+1}|Z_t, S_t, A_t)$ and the agent receives the reward R_{t+1} (given Z_t and S_{t+1}). Finally, the next input Z_{t+1} is generated using the transition function $P_z(Z_{t+1}|Z_t)$. Under this formulation, the next

[1] https://github.com/GRAAL-Research/Optimal_discounting_IDMDP

input only depends on the current input, but P_z can be extended such that it depends on all inputs previously observed in the trajectory (Mao et al., 2018).

Remark 1 (Augmented MDP)**.** *Given an IDMDP* $(\mathcal{S}, \mathcal{Z}, \mathcal{A}, P_s, P_z, R, \gamma)$*, one can define an augmented state space* $\bar{\mathcal{S}} = \mathcal{Z} \times \mathcal{S}$ *to parametrize a standard MDP* $(\mathcal{S}, \mathcal{A}, \bar{P}, \bar{R})$ *with* $\bar{P}(\bar{s}'|\bar{s}, a) = P_s(s'|s, a)P_z(z'|z)$ *and* $\bar{R}(\bar{s}, a) = R(s, z, a)$ *for inputs* $z, z' \in \mathcal{Z}$*, states* $s, s' \in \mathcal{S}$*, augmented states* $\bar{s}, \bar{s}' \in \bar{\mathcal{S}}$*, and action* $a \in \mathcal{A}$*. We refer to the resulting MDP as the* augmented MDP.

Model-Based Offline RL In model-based offline RL, the learning agent has access to a dataset $\mathcal{D} = \left\{\left(S_i, A_i, R_i, S_{i+1}, d_i\right)\right\}_{i=1}^{N}$ of N transition samples, where d_i is a boolean flag indicating whether the episode terminated after the transition. The dataset is typically collected using a mixture of behaviour policies such that the dataset is being sampled over what is called the behavioural distribution (Yu et al., 2020). The agent learns an approximate model of the environment $\widehat{M}$ by learning the dynamics $\widehat{P}(S_{t+1}|S_t, A_t)$ and $\widehat{R}(S_t, A_t)$ from the dataset. By performing synthetic rollouts, the agent aims to find $\pi^{\star}_{\widehat{M},\gamma}$, i.e. an optimal policy on $\widehat{M}$ for a discount factor γ.

The IDMDP formulation offers several advantages for offline model-based RL in applied (industrial) settings. Since inputs often evolve in ways that differ significantly from states, specialized modelling teams are justified in estimating $\widehat{P_z}$, leveraging domain expertise (e.g., pandemic or population evolution models). The IDMDP formulation also helps to reduce the variance when estimating values of actions, since it isolates the impact of the agent on the state space (Mao et al., 2018). Finally, inputs are often dependent on highly stochastic processes, which are challenging to model. This motivates the need to better understand the impacts of an imperfect input-model, and its ties to the optimal planning horizon in offline model-based IDMDPs.

3 Shallow planning in IDMDPs

In this section, we aim to caracterize the impacts of shallow planning in the offline IDMDP setting given an imperfect input-model.

Blackwell discount factor It is commonly assumed that a high discount factor (longer planning horizon) should lead to a better policy since it gives more information to the agent about the future impact of their actions. However, even with an infinite amount of data where one could have a perfect model of the MDP M, i.e. $\widehat{M} = M$, this is not always the case. It has been shown that there always exists a discount factor γ_{Bw} such that increasing the discount factor further does not result in a better policy (Grand-Clément & Petrik, 2024). Formally, for any $\gamma \geqslant \gamma_{\text{Bw}}$, we have $V^{\pi^{\star}_{M,\gamma}}_{M,\gamma} = V^{\pi^{\star}_{M,\gamma_{\text{Bw}}}}_{M,\gamma}$ when $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$. We refer to γ_{Bw} as the *Blackwell discount factor*. In other words, some MDPs might not require temporal tradeoffs (low γ_{Bw}), such that a low discount factor γ can lead to optimal behaviour even under a long-term objective. Therefore, any discount factor chosen above γ_{Bw} when data is limited and $\widehat{M}$ is imperfect only cumulates variance in the estimations of state values, resulting into poor generalization, commonly referred to as the planning loss (Lefebvre & Durand, 2025). The concept of Blackwell discount factor naturally extends to IDMDPs through their connections to MDPs (Remark 1). Hence, the Blackwell discount factor of an IDMDP is the Blackwell discount factor of its corresponding augmented MDP.

Planning loss A model-based RL agent aims to find an optimal policy on the approximate MDP $\widehat{M} \approx M$. When using a discount factor $\gamma < \gamma_{\text{Bw}}$ in such setting, the optimal policy is subject to a planning loss on the true environment M (Jiang et al., 2015):

$$\|V^{\pi^{\star}_{M,\gamma_{\text{Bw}}}}_{M,\gamma_{\text{Bw}}} - V^{\pi^{\star}_{\widehat{M},\gamma}}_{M,\gamma_{\text{Bw}}}\|_{\infty}. \leqslant \underbrace{\|V^{\pi^{\star}_{M,\gamma_{\text{Bw}}}}_{M,\gamma_{\text{Bw}}} - V^{\pi^{\star}_{M,\gamma_{\text{Bw}}}}_{M,\gamma}\|_{\infty}}_{\text{bias}} + \underbrace{\|V^{\pi^{\star}_{M,\gamma}}_{M,\gamma} - V^{\pi^{\star}_{\widehat{M},\gamma}}_{M,\gamma}\|_{\infty}}_{\text{variance}}. \tag{1}$$

The *bias* captures the loss in value function when using a policy that is optimal under the Blackwell discount factor, evaluated on a shallow horizon. The *variance* captures the impact of optimizing a policy under an approximate model $\widehat{M}$.

Several upper bounds on the planning loss exist in the literature for the MDP setting Jiang et al. (2015; 2016); Lefebvre & Durand (2025). In this work, we provide the first bounds the planning loss in IDMDPs by focusing on the distinctive feature of IDMDPs, i.e. the agent-independent inputs.

Assumption 1 (Exo-MDP)**.** *We therefore consider IDMDPs where P_s and R are known, i.e. $\widehat{P}_s = P_s$ and $\widehat{R} = R$, and focus on the impact of the approximate $\widehat{P}_z \approx P_z$. This is also known as the Exo-MDP setting (Sinclair et al., 2023).*

Inspired by the analysis of block contextual MDPs (Sodhani et al., 2022), we introduce a distance to measure how the value of a state changes depending on the input.

Definition 1 (Input metric)**.** *Let $z_i, z_j \in \mathcal{Z}$ denote two inputs from an IDMDP. Let $\bar{\mathcal{S}}$, $\bar{R}$, and $\bar{P}$ respectively denote the state space, reward function, and transition function in the corresponding augmented MDP (Remark 1). Let $s \in \mathcal{S}$ denote a state from the IDMDP and let $\bar{s}_i = (z_i, s)$, $\bar{s}_j = (z_j, s)$ denote the associated augmented states. We define the following distances for a given discount factor γ and policy $\pi : \mathcal{Z} \times \mathcal{S} \mapsto \mathcal{A}$:*

$$d^{\pi}_{\textit{states},\gamma}(\bar{s}_i, \bar{s}_j) := \left[\left| \bar{R}^{\pi}(\bar{s}_i) - \bar{R}^{\pi}(\bar{s}_j) \right| + \gamma W_1(d^{\pi}_{\textit{states},\gamma}) \left(\bar{P}^{\pi}(\bar{s}_i), \bar{P}^{\pi}(\bar{s}_j) \right) \right] \tag{2}$$

$$d^{\pi}_{\textit{input},\gamma}(z_i, z_j) := \max_{s \in \mathcal{S}} d^{\pi}_{\textit{states},\gamma}((z_i, s), (z_j, s)), \tag{3}$$

where W_1 is the Wasserstein distance, $\bar{R}^{\pi}(\bar{s}) := \sum_a \pi(a|\bar{s}) \bar{R}(\bar{s}, a)$ and $\bar{P}^{\pi}(\bar{s}) := \sum_a \pi(a|\bar{s}) \sum_{\bar{s}' \in C} \bar{P}(\bar{s}'|\bar{s}, a) \quad \forall C \in \bar{\mathcal{S}}_{E^{\pi}}$, with $\bar{\mathcal{S}}_{E^{\pi}}$ denoting the set of π-bisimilar groups of augmented-states (Castro, 2020).

Intuitively, if two inputs z_i and z_j lead to the same next inputs and next states while having leading to similar rewards, these inputs will be similar under this metric, i.e. their distance will be small. This distance captures the underlying dynamics in a succinct way and considers the worst case over all states. When we refer to the structure in the input space, we refer to the structure imposed by the input metric. Using Definition 1, we can bound the impact of inputs on state-values.

Lemma 1 (State-value difference under two inputs)**.** *Let $z_i, z_j \in \mathcal{Z}$ denote two inputs from an IDMDP M. For any state $s \in \mathcal{S}$, policy $\pi : \mathcal{Z} \times \mathcal{S} \mapsto \mathcal{A}$, and discount factor γ:*

$$|V^{\pi}_{M,\gamma}(s, z_i) - V^{\pi}_{M,\gamma}(s, z_j)| \leqslant d^{\pi}_{\textit{input},\gamma}(z_i, z_j). \tag{4}$$

This results from a direct application of Theorem 3 from Castro (2020) (see Supp. Sec. 8). Lemma 1 indicates that if two inputs are similar under the input metric (Definition 1), the values of any state augmented with these inputs should be close. Lemma 1 (which also holds for augmented states) generalizes the value function variation from Jiang et al. (2016) to any policy and any pair of states.

Lemma 2 (Variance)**.** *Consider optimal policies (for a given discount factor γ) computed on an IDMDP M (with input transition function P_z) and on an approximate model $\widehat{M}$ (with approximate input transition function $\widehat{P}_z$). The difference in their value-functions evaluated on M is bounded by:*

$$\|V^{\pi^{\star}_{M,\gamma}}_{M,\gamma} - V^{\pi^{\star}_{\widehat{M},\gamma}}_{M,\gamma}\|_{\infty} \leqslant \frac{\gamma}{(1-\gamma)} \max_{z} \|\widehat{P}_z(\cdot|z) - P_z(\cdot|z)\|_1 \max_{z_i, z_j \in \mathcal{Z}} \max_{\pi : \mathcal{Z} \times \mathcal{S} \mapsto \mathcal{A}} d^{\pi}_{\textit{input},\gamma}(z_i, z_j). \tag{5}$$

Roughly speaking, if $\widehat{P}_z$ is a *bad* approximation of P_z for inputs that have a strong impact on the underlying dynamics (rewards and state transitions), the optimal policy on $\widehat{M}$ will not generalize well on the real environment M. Lemma 2 provides further insight into the behaviour of the variance, complementing prior results (Jiang et al., 2015; Lefebvre & Durand, 2025) by emphasizing the impact of non-controllable variables (inputs) and their modelling error. One should also observe that the bound goes to 0 if the agent is myopic ($\gamma = 0$). This indicates that a way to reduce the impact of an imperfect model on the inputs is to reduce the planning horizon. Using Lemma 1 from Jiang et al. (2015) and Lemma 2 above, we can upper-bound the planning loss (Equation 1, see Supp. Sec. 10).

Theorem 1 (Planning loss)**.** *Given an Exo-MDP M using P_z and its approximation $\widehat{M}$ using $\widehat{P}_z$, learning a policy on $\widehat{M}$ using discount factor $\gamma \leqslant \gamma_{Bw}$, the planning loss is bounded by:*

$$\|V^{\pi^\star_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}} - V^{\pi^\star_{\widehat{M},\gamma}}_{M,\gamma_{Bw}}\|_\infty \leqslant \frac{\gamma_{Bw}-\gamma}{(1-\gamma_{Bw})(1-\gamma)} R_{max} + \frac{\gamma}{(1-\gamma)} \max_z \|\widehat{P}_z(\cdot|z) - P_z(\cdot|z)\|_1 \max_{z_i,z_j\in\mathcal{Z}} \max_{\pi:\mathcal{Z}\times\mathcal{S}\mapsto\mathcal{A}} d^\pi_{input,\gamma}(z_i,z_j).$$

As expected, reducing the planning horizon (lowering the discount factor) increases the bias (1st term) while decreasing the variance (2nd term). The variance also decreases as the quality of the input model approximate improves or as input impact similarity increases (input metric decreases). One can use this result to help select the discount factor in practice. For instance, when modelling a complex input space with few data points, one can expect high variance justifying a lower planning horizon to mitigate it. By analyzing the planning loss under Exo-MDPs, Theorem 1 highlights how highly stochastic input transitions and an approximate input-model impacts the planning loss.

4 Input-dependent planning

It is natural to believe that the knowledge of the agent may not be uniform over the input space $\mathcal{Z}$. Indeed, it very possible that the approximate input dynamics $\widehat{P}_z$ may be closer to the true dynamics P_z in some regions of the input space than others. Based on Theorem 1, a better knowledge of the input dynamics would warrant a longer planning horizon. We refer to the discount factor that minimizes the planning loss for a given input z as its *optimal input-dependent discount factor*:

$$\gamma^\star(z) := \operatorname*{argmin}_{\gamma\in[0,\gamma_{\text{Bw}}]} \|V^{\pi^\star_{M,\gamma_{\text{Bw}}}}_{M,\gamma_{\text{Bw}}}(\cdot,z) - V^{\pi^\star_{\widehat{M},\gamma}}_{M,\gamma_{\text{Bw}}}(\cdot,z)\|_\infty \tag{6}$$

To understand the behavior of the optimal input-dependent discount factor over the input space, we introduce the following result on the planning loss difference between two inputs.

Lemma 3 (Input-wise planning loss difference)**.** *Given an IDMDP M and its approximation $\widehat{M}$, and let $z_i, z_j \in \mathcal{Z}$ denote two inputs. For any discount factor $\gamma \in [0, \gamma_{Bw}]$ we have:*

$$|f_{z_i}(\gamma) - f_{z_j}(\gamma)| \leqslant 2 \max_{\pi:\mathcal{Z}\times\mathcal{S}\mapsto\mathcal{A}} d^\pi_{input,\gamma_{Bw}}(z_i,z_j),$$

where $f_z(\gamma) = \|V^{\pi^\star_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}}(\cdot,z) - V^{\pi^\star_{\widehat{M},\gamma}}_{M,\gamma_{Bw}}(\cdot,z)\|_\infty$ denotes the planning loss at input z.

The proof essentially relies on the application of the triangle inequality combined with the fact that a maximum is infinity-norm Lipschitz (see Supp. Sec. 11). From prior results Jiang et al. (2015); Hu et al. (2022) we see that the planning loss tend to be U-shaped. Following this, we will use the following assumption, which enables Lemma 3 to bound the difference between optimal input-dependent discount factors.

Assumption 2 (Convexity of the planning loss)**.** *For any discount factors $\gamma, \gamma_0 \in [0, \gamma_{Bw}]$:*

$$f_z(\gamma) \geqslant f_z(\gamma_0) + \langle \nabla f_z(\gamma_0), \gamma - \gamma_0\rangle + \frac{\mu}{2}|\gamma-\gamma_0|^2, \tag{7}$$

where $f_z(\gamma) = \|V^{\pi^\star_{M,\gamma_{Bw}}}_{M,\gamma_{Bw}}(\cdot,z) - V^{\pi^\star_{\widehat{M},\gamma}}_{M,\gamma_{Bw}}(\cdot,z)\|_\infty$ denotes the planning loss at input z.

Theorem 2 (Optimal input-dependent discount factor)**.** *Given an IDMDP M and its approximation $\widehat{M}$, and let $z_i, z_j \in \mathcal{Z}$ denote two inputs. Assuming that the planning loss is μ-strongly convex, we have:*

$$|\gamma^\star(z_i) - \gamma^\star(z_j)| \leqslant \sqrt{\frac{8\max_{\pi:\mathcal{Z}\times\mathcal{S}\mapsto\mathcal{A}} d^\pi_{input,\gamma_{Bw}}(z_i,z_j)}{\mu}}. \tag{8}$$

The proof essentially relies on the triangle inequality (see Supp. Sec. 12). Theorem 2 reinforces the intuition that similar inputs should have similar optimal planning horizons. It is important to note that the distance metric is under the Blackwell discount factor which means that the distance between two inputs is measured on their long term differences in dynamics. If two inputs are just temporally near, they might still remain distant under the distance metric, s.t. that their optimal planning horizon could differ. Another important part of the bound is the convexity μ. Strong convexity assumes that the curvature (second derivative) is always higher than some threshold μ. A high μ indicates that the planning loss is very sensitive to the discount factor, where two similar discount factors could lead to very different planning losses and vice versa. Therefore, depending on μ and the local structure around an input, the optimal planning horizon will tend to be similar for groups of long-term similar inputs. This result motivates further investigation of input-dependent planning horizons.

5 Experiments

We now conduct experiments to support our bound on the planning loss (Theorem 1) and our result on the smoothness of the optimal input-dependent discount factor (Theorem 2).

5.1 Ring IDMDP

Validating Theorem 1 requires an environment for which we can control the different quantities appearing in the bound and find the true optimal state-value (for γ_{Bw}). We therefore consider the Ring MDP (Jiang et al., 2016) setting. *Ring*(N, p) is an MDP with 2 actions, $\mathcal{A} = \{1, 2\}$, and N states arranged in a ring. Action 1 moves the agent clockwise, while action 2 moves the agent counter-clockwise. For each pair of non-adjacent states (s_i, s_j), we add an edge to s_j from s_i given action a with probability p for each action $a \in \mathcal{A}$. The transition probabilities over all edges are then uniformely sampled from $[0, 1]$ and normalized. Similarly, the mean rewards are assigned to every state-action pairs by uniformly sampling from $[0, 1]$ s.t. $R_{\text{max}} = 1$.

From a Ring MDP, we generate a Ring IDMDP with input space $\mathcal{Z} = \{0, 1\}$. For input $z = 0$, state transition probabilities and expected rewards are given by $P_s(s'|0, s, a)$ and $R(0, s, a)$. For input $z = 1$, the impact of actions are blended (with scaling parameter $\alpha \in [0, 1]$):

$$P(s'|1, s, a; \alpha) = (1 - \alpha)\, P(s'|0, s, a) + \alpha\, P(s'|0, s, \bar{a})$$
$$R(1, s, a; \alpha) = (1 - \alpha)\, R(0, s, a) + \alpha \left[-\big|R(0, s, a,)\big| \right],$$

where $\bar{a}$ denotes the anti-action, i.e. $\bar{a} = 1$ if $a = 0$, else $\bar{a} = 0$. For $\alpha = 0$, inputs have no impact. For $\alpha = 1$, input $z = 1$ inverts the dynamics. Inputs evolve according to $P_z(z'|z) = 0.5$ for all z, z'.

Approximate Ring IDMDPs are generated by sampling at random imperfect kernels $\widehat{P}_z$ using rejection sampling to enforce constraints on the total model-error $E = \sum_z \|P_z(\cdot|z) - \widehat{P}_z(\cdot|z)\|_1$. We evaluate the impact of the planning horizon on the normalized planning loss (with $\gamma_{\text{Bw}} = 0.99$):

$$\max_{s \in \mathcal{S}} \left(V^{\pi^\star_{M,\gamma_{\text{Bw}}}}_{M,\gamma_{\text{Bw}}}(s) - V^{\pi^\star_{\widehat{M},\gamma}}_{M,\gamma_{\text{Bw}}}(s) \right) / V^{\pi^\star_{M,\gamma_{\text{Bw}}}}_{M,\gamma_{\text{Bw}}}(s). \tag{9}$$

To illustrate the impact of input similarity on Theorem 1, we sample 10^4 approximate models with $E > 1.75$[2]. We average the normalized planning loss over all approximate Ring IDMDPs for each $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$; recall that higher α means more distance between inputs, i.e. more impact of inputs on P and R. To illustrate the impact of model-error on Theorem 1, we sample 10^4 approximate models for each total model-error range $E \in \{[0, 1.5], [1.5, 1.75], [1.75, 2]\}$. For each range, we average the normalized planning loss over all models in this range (using a fixed $\alpha = 1$).

Figure 1 shows that the planning loss is heavily impacted by the input-distance and model-error. Increasing α compounds with model-error, which increases the variance, making shallow planning

[2]The maximum total model-error on the P_z kernel is 2.

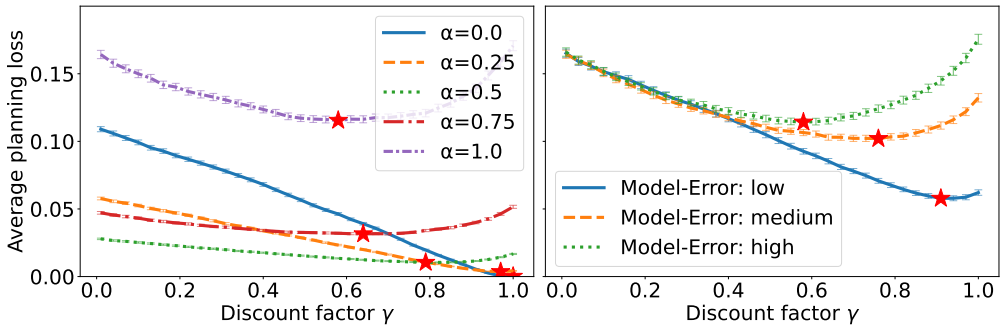

Figure 1: Average normalized planning loss (Equation 9) for different input influence (left) and total model-error (right) given the discount factor. The optimal planning horizon is marked by a star.

(lower discount factors) beneficial (left). For a fixed α, the optimal planning horizon increases as model-error decreases (right). These results support Theorem 1.

5.2 Input-driven CartPole

We investigate our theoretical results under a high-dimensional setting, i.e a variant of the CartPole-v1 environment (Towers et al., 2024). In CartPole, the agent uses two discrete actions (left or right) to balance a pole on a moving cart. The state contains the cart's position and velocity, along with the pole's angle and velocity.

Reward zones We define the input space $\mathcal{Z} = \{0, 1, 2, 3\}$, which controls the location of a *pay zone*. The environment contains two zones: a *safe zone* in the middle, where the cart starts, and a pay zone, whose position depends on the input. For inputs 0 and 2, the pay zone is on the left; for 1 and 3, it is on the right. The reward is $+1.25$ in the pay zone, $+0.5$ in the safe zone, and -0.25 elsewhere. Inputs evolve as follows: $P_z(0|0) = P_z(1|1) = 1$; $P_z(2|3) = P_z(3|2) = 0.9$; $P_z(2|2) = P_z(3|3) = 0.1$. Inputs 2 and 3 are thus similar under Definition 1. Under input 2 or 3, planning far ahead with an approximate model is often suboptimal; the agent should favour short-term rewards in the safe zone. The trajectory ends when the pole falls or after 500 steps. This setting reflects many real-world problems where long-term dynamics are uncertain. For example, a retailer may offer discounts assuming stable purchasing behaviour, only to see it shift unexpectedly, reducing profits.

We consider Model-based Offline Policy Optimization (MOPO) agents (Yu et al., 2020) based on discrete Soft Actor-Critic (Christodoulou, 2019). We train 10 agents with each discount factor $\gamma \in \{0.92, 0.94, 0.96, 0.98, 0.99\}$ on an offline *medium-expert* dataset comprised of 10^5 transitions from several policy (Fu et al., 2020). We consider two expert (optimal) policies (using $\gamma = 0.98$) with baseline parameters (Raffin et al., 2021) and a random policy: oracle expert has acess to P_z while the approximate expert assumes static inputs, i.e. $\widehat{P}_z(z'|z) = 1$ if $z' = z$, else 0. MOPO agents optimize their policies on the dataset using the learnt dynamics $\widehat{P}_s$ (assuming the rewards are known using S_{t+1}), using here again an approximate input model $\widehat{P}_z$ which assumes static inputs (worst case scenario). We evaluate the resulting 50 models on 100 seeds of the environment[3].

Figure 2 confirms that the optimal discount factor changes according to the input. When the pay zone position is well-known (inputs 0 and 1), a longer planning horizon (larger γ) enables the agent to tradeoff immediate rewards in order to move to the high-paying pay zone. On the other hand, when the pay zone position cannot be well-predicted, long-term planning tends to overfit to the dataset (approximate model), which can result in catastrophic performance. These results support Theorem 1. They also shows the benefits of adapting the planning horizon to the local model-error, supporting Theorem 2 and motivating further work on input-dependent discount factors.

[3]See Supp. Sec. 13 for hyperparameters.

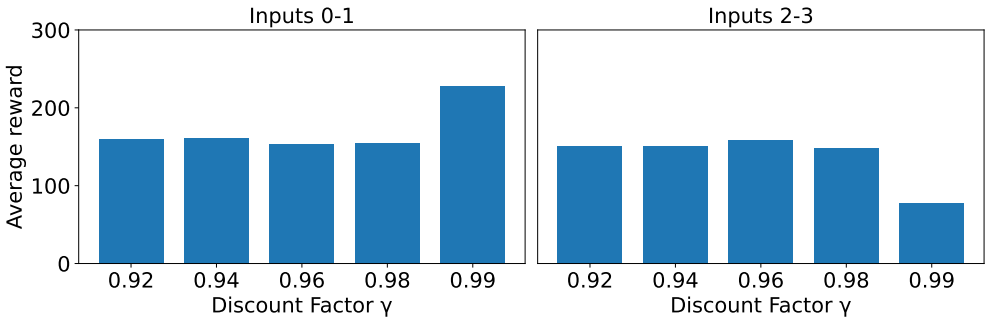

Figure 2: Average reward obtained by the 10 agents trained with each γ over 100 evaluations depending on the starting input Z_0.

6 Related Work

Bias-variance tradeoff with shallow planning The bias-variance tradeoff arising from shallow planning on an imperfect model has been well studied in (PO)MDPs (Jiang et al., 2015; 2016; Lefebvre & Durand, 2025). In IDMDPs, this translates into a focus on how the error in the approximate state model $\widehat{P}_s$ impacts the planning loss. In this work, we rather focus on the impacts of an imperfect input model $\widehat{P}_z$, knowing that inputs (unlike states) are not controlled by the agent.

State/Input-dependent discounting The question of using a different discount factor on different part of the state/input space has received low attention in RL. In MDPs, it has been shown that the optimality criterion when using a state-specific discount factor is well-defined and has a unique solution (Wei & Guo, 2011). Despite good theoretical foundations, tuning a state-specific discount factor is highly non-trivial (Yoshida et al., 2013). An alternative recent avenue suggests planning with a uniform horizon on locally-regularized transitions according to the uncertainty of state-action pairs (Rathnam et al., 2024), which does not apply to non-controlable variables (e.g. inputs). To our knowledge, there are currently no results on input-dependent planning in IDMDPs.

7 Conclusion

This work provides the first bias-variance tradeoff analysis in offline learning under Input-Driven MDPs. Focusing on the error arrising from input modelling, we provide new insights connecting input structure to the state-value function (Lemma 1). This leads to a novel bound on the variance (Lemma 2), which we use to obtain the first existing bound on the planning loss for Exo-MDPs (Theorem 1). These insights indicate that the discount factor impacts generalization, as expected from existing results in MDPs. We further investigate the optimal discount factor *per input*. We connect the planning loss at different inputs to the input structure (Lemma 3), allowing to control the variation of optimal input-dependent discount factors over the input space using the input structure (Theorem 2). These results complements the work on state-dependent discount factors.

Limitations and future work This work focuses on the impact of approximating input transitions. In reality, the model of state transitions is also usually imperfect, s.t. the discount factor generalization properties depend on the compounding error of both the input and state models. Our univariate analysis therefore offers only a partial picture, motivating future work that would combine our analysis with existing work (Jiang et al., 2015; 2016; Lefebvre & Durand, 2025) on the bias-variance tradeoff in MDPs. Through the augmented MDP (Remark 1), we also clearly see that our results extend trivially to the MDP framework, thus providing a novel pathway to tackle state-dependent planning under the partial observability induced by uncontrollable input variables.

References

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, 2020.

Ron Amit, Ron Meir, and Kamil Ciosek. Discount factor as a regularizer in reinforcement learning. In *International Conference on Machine Learning*, 2020.

Arthur Argenson and Gabriel Dulac-Arnold. Model-based offline planning. In *International Conference on Learning Representations*, 2021.

Loris Cannelli, Giuseppe Nuti, Marzio Sala, and Oleg Szehr. Hedging using reinforcement learning: Contextual k-armed bandit versus Q-learning. *The Journal of Finance and Data Science*, 2023.

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic markov decision processes. In *AAAI Conference on Artificial Intelligence*, 2020.

Petros Christodoulou. Soft actor-critic for discrete action settings. *arXiv:1910.07207*, 2019.

Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. An empirical investigation of the challenges of real-world reinforcement learning. *Machine Learning*, 2021.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv:2004.07219*, 2020.

Dibya Ghosh, Jad Rahme, Aviral Kumar, Amy Zhang, Ryan P Adams, and Sergey Levine. Why generalization in RL is difficult: Epistemic POMDPs and implicit partial observability. *Advances in Neural Information Processing Systems*, 2021.

Julien Grand-Clément and Marek Petrik. Reducing blackwell and average optimality to discounted MDPs via the blackwell discount factor. *Advances in Neural Information Processing Systems*, 2024.

Hao Hu, Yiqin Yang, Qianchuan Zhao, and Chongjie Zhang. On the role of discount factor in offline reinforcement learning. In *International Conference on Machine Learning*, 2022.

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. In *International Conference on Autonomous Agents and Multiagent Systems*, 2015.

Nan Jiang, Satinder Singh, and Ambuj Tewari. On structural properties of MDPs that bound loss due to shallow planning. In *International Joint Conference on Artificial Intelligence*, 2016.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 2020.

Randy Lefebvre and Audrey Durand. On shallow planning under partial observability. In *AAAI Conference on Artificial Intelligence*, 2025.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *Perspectives on Open Problems*, 2020.

Yi Liu and Lihong Li. A map of bandits for e-commerce. *arXiv:2107.00680*, 2021.

Hongzi Mao, Shaileshh Bojja Venkatakrishnan, Malte Schwarzkopf, and Mohammad Alizadeh. Variance reduction for reinforcement learning in input-driven environments. *arXiv:1807.02264*, 2018.

Volodymyr Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 2015.

Haoyi Niu, Yiwen Qiu, Ming Li, Guyue Zhou, Jianming Hu, Xianyuan Zhan, et al. When to trust your simulator: Dynamics-aware hybrid offline-and-online reinforcement learning. *Advances in Neural Information Processing Systems*, 2022.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 2021.

Sarah Rathnam, Sonali Parbhoo, Siddharth Swaroop, Weiwei Pan, Susan A Murphy, and Finale Doshi-Velez. Rethinking discount regularization: New interpretations, unintended consequences, and solutions for regularization in reinforcement learning. *Journal of Machine Learning Research*, 2024.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016.

Sean R Sinclair, Felipe Vieira Frujeri, Ching-An Cheng, Luke Marshall, Hugo De Oliveira Barbalho, Jingling Li, Jennifer Neville, Ishai Menache, and Adith Swaminathan. Hindsight learning for mdps with exogenous inputs. In *International Conference on Machine Learning*, 2023.

Shagun Sodhani, Franziska Meier, Joelle Pineau, and Amy Zhang. Block contextual mdps for continual learning. In *Learning for Dynamics and Control Conference*, 2022.

Aaron Sonabend, Junwei Lu, Leo Anthony Celi, Tianxi Cai, and Peter Szolovits. Expert-supervised reinforcement learning for offline policy learning and evaluation. *Advances in Neural Information Processing Systems*, 2020.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulao, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv:2407.17032*, 2024.

Qingda Wei and Xianping Guo. Markov decision processes with state-dependent discount factors and unbounded rewards/costs. *Operations Research Letters*, 2011.

Naoto Yoshida, Eiji Uchibe, and Kenji Doya. Reinforcement learning with state-dependent discount factor. In *IEEE Joint International Conference on Development and Learning and Epigenetic Robotics*, 2013.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y. Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 2020.

Supplementary Materials

The following content was not necessarily subject to peer review.

8 Proof of Lemma 1

We first leverage Remark 1 to turn the IDMDP into its augmented MDP, and then use Theorem 3 from Castro (2020), which requires the following result:

Theorem 3 (Theorem 2 from Castro (2020))**.** *Given the MDP M, two states $s, t \in \mathcal{S}$ and a pseudo-metric d on $\mathcal{S}$, define the operator $F^\pi : M \to M$ by*

$$F^\pi(d)(s,t) = |R^\pi(s) - R^\pi(t)| + \gamma\, W_1\big(d\big)(P^\pi(s), P^\pi(t)).$$

Then F^π has a least fixed point $\tilde{d}^\pi$, and $\tilde{d}^\pi$ is a π-bisimulation metric. W_1 is the wasserstein distance, $\bar{R}^\pi(\bar{s}) := \sum_a \pi(a|\bar{s})\bar{R}(\bar{s},a)$ and $\bar{P}^\pi(\bar{s}) = \sum_a \pi(a|\bar{s}) \sum_{\bar{s}'\in C} \bar{P}(\bar{s}'|\bar{s},a) \quad \forall C \in \bar{\mathcal{S}}_{E^\pi}$ where $\bar{\mathcal{S}}_{E^\pi}$ denotes every group of (augmented) states which are π-bisimilar (Castro, 2020).

By using the definition of a fixed point $\tilde{d}^\pi = F^\pi(\tilde{d}^\pi)$, we recover the distance metric:

$$d^\pi_{\text{states}}(\bar{s}, \bar{s}') := \Big[\big|R^\pi(\bar{s}) - R^\pi(\bar{s}')\big| + \gamma W(d^\pi_{\text{states}})\Big(P^\pi(\bar{s}), P^\pi(\bar{s}')\Big)\Big]. \tag{10}$$

By realizing that $(s, z_i), (s, z_j) \in \bar{\mathcal{S}}$, we can use the following theorem:

Theorem 4 (Theorem 3 from Castro (2020))**.** *For any two states $s, t \in \mathcal{S}$ in an MDP,*

$$|V^\pi(s) - V^\pi(t)| \leqslant d^\pi_{states}(s,t). \tag{11}$$

We therefore have:

$$\begin{aligned}
|V^\pi_{M,\gamma}((s,z_i)) - V^\pi_{M,\gamma}((s,z_j))| &\leqslant d^\pi_{\text{states}}((s,z_i),(s,z_j)) \\
&\leqslant \max_s d^\pi_{\text{states}}((s,z_i),(s,z_j)) \\
&= d^\pi_{\text{inputs},\gamma}(z_i, z_j),
\end{aligned}$$

which is the desired result.

9 Proof of Lemma 2

Our proof relies on the following existing results:

Lemma 4 (Lemma 3 from Jiang et al. (2015))**.** *For any mdp $\widehat{M} = (\mathcal{S}, \mathcal{A}, R, \widehat{P}, \gamma)$*

$$\|V^{\pi^*_{M,\gamma}}_{M,\gamma} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma}\|_\infty \leqslant 2 \max_{\pi:\mathcal{S}\mapsto\mathcal{A}} \|V^\pi_{M,\gamma} - V^\pi_{\widehat{M},\gamma}\|_\infty. \tag{12}$$

Lemma 5 (Lemma 4 from Jiang et al. (2015))**.** *For any mdp $\widehat{M} = (\mathcal{S}, \mathcal{A}, R, \widehat{P}, \gamma)$, $\forall \pi : \mathcal{S} \mapsto \mathcal{A}$,*

$$\|Q^\pi_{M,\gamma} - Q^\pi_{\widehat{M},\gamma}\|_\infty \leqslant \frac{1}{1-\gamma} \max_{s\in\mathcal{S}, a\in\mathcal{A}} \Big|R(s,a) + \gamma\langle \widehat{P}(\cdot|s,a), V^\pi_{M,\gamma}\rangle - Q^\pi_{M,\gamma}(s,a)\Big|. \tag{13}$$

From the IDMDP tuple $(\mathcal{S}, \mathcal{Z}, \mathcal{A}, R, P_s, P_z, \gamma)$, we define the equivalent augmented MDP (Remark 1) with $M = (\bar{\mathcal{S}} = \mathcal{S} \times \mathcal{Z}, \mathcal{A}, R, P, \gamma)$ with $P(\bar{S}_{t+1}|\bar{S}_t, A_t) = P_s(S_{t+1}|S_t, A_t)P_z(z_{t+1}|z_t)$.

The approximate augmented MDP is $\widehat{M} = (\bar{\mathcal{S}}, \mathcal{A}, R, \widehat{P}, \gamma)$ with $\widehat{P}(\bar{S}_{t+1}|\bar{S}_t, A_t) = P_s(S_{t+1}|S_t, A_t)\widehat{P}_z(z_{t+1}|z_t)$. We can therefore use Lemma 5 and Lemma 4 to obtain:

$$\begin{aligned}
\|V^{\pi^*_{M,\gamma}}_{M,\gamma} - V^{\pi^*_{\widehat{M},\gamma}}_{M,\gamma}\|_\infty &\leqslant 2 \max_{\pi:\bar{\mathcal{S}}\mapsto\mathcal{A}} \|V^\pi_{M,\gamma} - V^\pi_{\widehat{M},\gamma}\|_\infty && (14)\\
&\leqslant 2 \max_{\pi:\bar{\mathcal{S}}\mapsto\mathcal{A}} \|Q^\pi_{M,\gamma} - Q^\pi_{\widehat{M},\gamma}\|_\infty && (15)\\
&\leqslant \frac{2}{1-\gamma} \max_{\substack{\bar{s}\in\bar{\mathcal{S}}, a\in\mathcal{A},\\ \pi:\bar{\mathcal{S}}\mapsto\mathcal{A}}} \left| R(\bar{s}, a) + \gamma\langle \widehat{P}(\cdot|\bar{s}, a), V^\pi_{M,\gamma}\rangle - Q^\pi_{M,\gamma}(\bar{s}, a)\right| && (16)\\
&= \frac{2\gamma}{1-\gamma} \max_{\substack{\bar{s}\in\bar{\mathcal{S}}, a\in\mathcal{A},\\ \pi:\bar{\mathcal{S}}\mapsto\mathcal{A}}} \left|\langle \widehat{P}(\cdot|\bar{s}, a) - P(\cdot|\bar{s}, a), V^\pi_{M,\gamma}\rangle\right|, && (17)
\end{aligned}$$

where the last line is obtain using the Q-value definition. From now on, we will refer to the state value $V^\pi_{M,\gamma}(\bar{s})$ (with $\bar{s} \in \bar{\mathcal{S}}$) as $V^\pi(s, z)$ (with $s, z \in \mathcal{S} \times \mathcal{Z}$) to alleviate the notation while expliciting the inputs and states underlying the augmented state. We will now focus our attention on the interior of the absolute value. First, we define a quantity:

$$\phi(s, \pi) = \frac{\max_z V^\pi(s, z) + \min_z V^\pi(s, z)}{2}, s \in \mathcal{S}, \pi : \bar{\mathcal{S}} \mapsto \mathcal{A}. \tag{18}$$

Under this quantity, we have the following equality:

$$\sum_{z'}\sum_{s'} P_s(s'|s, a, z)\widehat{P}_z(z'|z)\phi(s', \pi) = \sum_{z'}\sum_{s'} P_s(s'|s, a, z)P_z(z'|z)\phi(s', \pi). \tag{19}$$

Proof.

$$\begin{aligned}
&\sum_{z'}\sum_{s'} P_s(s'|s, a, z)\widehat{P}_z(z'|z)\phi(s', \pi) - \sum_{z'}\sum_{s'} P_s(s'|s, a, z)P_z(z'|z)\phi(s', \pi)\\
&= \sum_{z'}\sum_{s'} P_s(s'|s, a, z)\phi(s', \pi)(\widehat{P}_z(z'|z) - P_z(z'|z))\\
&= \sum_{s'} P_s(s'|s, a, z)\phi(s', \pi)\sum_{z'}(\widehat{P}_z(z'|z) - P_z(z'|z))\\
&= \sum_{s'} P_s(s'|s, a, z)\phi(s', \pi)(0)\\
&= 0
\end{aligned}$$

□

Therefore, for the interior of the absolute value, we have:

$$
\begin{aligned}
&\langle \widehat{P}(\cdot|\bar{s},a) - P(\cdot|\bar{s},a), V^{\pi}_{M,\gamma}\rangle \\
&\quad= \sum_{z'}\sum_{s'}\Big(P_s(s'|s,a,z)\widehat{P}_z(z'|z)V^{\pi}(s',z') - P_s(s'|s,a,z)P_z(z'|z)V^{\pi}(s',z')\Big) \\
&\quad= \sum_{z'}\sum_{s'} P_s(s'|s,a,z)\widehat{P}_z(z'|z)\left(V^{\pi}(s',z') - \phi(s',\pi)\right) \\
&\qquad\quad - P_s(s'|s,a,z)P_z(z'|z)\left(V^{\pi}(s',z') - \phi(s',\pi)\right) \\
&\quad= \sum_{z'}(\widehat{P}_z(z'|z) - P_z(z'|z))\sum_{s'} P_s(s'|s,a,z)\left(V^{\pi}(s',z') - \phi(s',\pi)\right) \\
&\quad\leqslant \|\widehat{P}_z(\cdot|z) - P_z(\cdot|z)\|_1 \max_{z'}\left|\sum_{s'} P_s(s'|s,a,z)\left(V^{\pi}(s',z') - \phi(s',\pi)\right)\right| \\
&\quad\leqslant \|\widehat{P}_z(\cdot|z) - P_z(\cdot|z)\|_1 \max_{z'}\left|\max_{s'}\left|V^{\pi}(s',z') - \phi(s',\pi)\right|\right| \\
&\quad= \|\widehat{P}_z(\cdot|z) - P_z(\cdot|z)\|_1 \max_{z'}\left|\max_{s'}\left|V^{\pi}(s',z') - \frac{\max_z V^{\pi}(s',z) + \min_z V^{\pi}(s',z)}{2}\right|\right| \\
&\quad\leqslant \frac{1}{2}\|\widehat{P}_z(\cdot|z) - P_z(\cdot|z)\|_1 \max_{s'}(\max_z V^{\pi}(s',z) - \min_z V^{\pi}(s',z)) \\
&\quad\leqslant \frac{1}{2}\|\widehat{P}_z(\cdot|z) - P_z(\cdot|z)\|_1 \max_{z_i,z_j} d^{\pi}_{\text{input},\gamma}(z_i,z_j).
\end{aligned}
$$

We use Holder's inequality on the first two inequalities, then substitute ϕ. We then realize that the upper bound between the magnitude of the difference in state-value and its middle point for any value of s is bounded by half the distance between the maximum and minimum values, and use Lemma 1. Plugging this into the initial bound (Equation 17), we get the desired result.

10 Proof Theorem 1

The goal is to bound the planning loss from equation 1.

$$
\|V^{\pi^{\star}_{M,\gamma_{\text{Bw}}}}_{M,\gamma_{\text{Bw}}} - V^{\pi^{\star}_{\widehat{M},\gamma}}_{M,\gamma_{\text{Bw}}}\|_{\infty}. \leqslant \underbrace{\|V^{\pi^{\star}_{M,\gamma_{\text{Bw}}}}_{M,\gamma_{\text{Bw}}} - V^{\pi^{\star}_{M,\gamma_{\text{Bw}}}}_{M,\gamma}\|_{\infty}}_{\text{bias}} + \underbrace{\|V^{\pi^{\star}_{M,\gamma}}_{M,\gamma} - V^{\pi^{\star}_{\widehat{M},\gamma}}_{M,\gamma}\|_{\infty}}_{\text{variance}}. \tag{20}
$$

We know from Jiang et al. (2015) that the bias is bounded:

Lemma 6 (Lemma 1 from Jiang et al. (2015))**.** *For any MDP M with rewards in $[0, R_{max}]$, $\forall \pi : \mathcal{S} \to \mathcal{A}$ and $\gamma \leqslant \gamma_{eval}$, $\forall s \in \mathcal{S}$,*

$$
V^{\pi^{\star}_{M,\gamma_{eval}}}_{M,\gamma_{eval}}(s) - V^{\pi^{\star}_{M,\gamma_{eval}}}_{M,\gamma}(s) \leqslant \frac{\gamma_{eval} - \gamma}{(1-\gamma_{eval})(1-\gamma)} R_{max} \tag{21}
$$

By replacing γ_{eval} by γ_{Bw}, we find the bias bound in Theorem 1. For the variance term, we upper bound it using Lemma 2 without any additional modifications. The final assembly of both result is given by:

$$
\begin{aligned}
\|V^{\pi^{\star}_{M,\gamma_{\text{Bw}}}}_{M,\gamma_{\text{Bw}}} - V^{\pi^{\star}_{\widehat{M},\gamma}}_{M,\gamma_{\text{Bw}}}\|_{\infty} \leqslant\ & \frac{\gamma_{\text{Bw}} - \gamma}{(1-\gamma_{\text{Bw}})(1-\gamma)} R_{\max} \\
&+ \frac{\gamma}{(1-\gamma)} \max_{z} \|\widehat{P}_z(\cdot|z) - P_z(\cdot|z)\|_1 \max_{z_i,z_j\in\mathcal{Z}} \max_{\pi:\mathcal{Z}\times\mathcal{S}\mapsto\mathcal{A}} d^{\pi}_{\text{input},\gamma}(z_i,z_j).
\end{aligned}
$$

Which is Theorem 1.

11 Proof of Lemma 3

Let $f_z(\gamma) = \|V^{\pi^\star_{M,\gamma_{\text{Bw}}}}_{M,\gamma_{\text{Bw}}}(\cdot, z) - V^{\pi^\star_{\widehat{M},\gamma}}_{M,\gamma_{\text{Bw}}}(\cdot, z)\|_\infty$ denote the planning loss given input z. Let us also define hortcuts to alleviate the notation:

$$V^{\pi^\star_{M,\gamma_{\text{Bw}}}}_{M,\gamma_{\text{Bw}}}(s, z) := V^\star(s, z)$$
$$V^{\pi^\star_{\widehat{M},\gamma}}_{M,\gamma_{\text{Bw}}}(s, z) := \widehat{V}(s, z).$$

The first denotes the value of state s at input z with the optimal policy on model M and discount factor γ_{Bw}. The seconds denotes the value of state s at input z of the optimal policy on approximate model $\widehat{M}$ and discount factor $\gamma < \gamma_{\text{Bw}}$ (when evaluated on true model M with discount factor γ_{Bw}). We can bound the difference in planning losses between two inputs z_i and z_j for a given factor γ:

$$\begin{aligned}
|f_{z_i}(\gamma) - f_{z_j}(\gamma)| &= \left|\max_s |V^\star(s, z_i) - \widehat{V}(s, z_i)| - \max_s |V^*(s, z_j) - \widehat{V}(s, z_j)|\right| \\
&\leqslant \left|\max_s \left||V^\star(s, z_i) - \widehat{V}(s, z_i)| - |V^\star(s, z_j) - \widehat{V}(s, z_j)|\right|\right| \\
&\leqslant \max_s \left|\left(V^\star(s, z_i) - \widehat{V}(s, z_i)\right) - \left(V^\star(s, z_j) - \widehat{V}(s, z_j)\right)\right| \\
&= \max_s \left|(V^\star(s, z) - V^\star(s, z_j)) + \left(\widehat{V}(s, z_j) - \widehat{V}(s, z_i)\right)\right| \\
&\leqslant \max_s |V^\star(s, z_i) - V^\star(s, z_j)| + \max_s \left|\widehat{V}(s, z_j) - \widehat{V}(s, z_i)\right| \\
&\leqslant 2 \max_{\pi:\bar{\mathcal{S}}\mapsto\mathcal{A}} d^\pi_{\text{input},\gamma_{\text{Bw}}}(z_i, z_j).
\end{aligned}$$

The first equality uses the fact that a maximum is infinity norm Lipschitz. We then use the triangle inequality first to obtain that the difference of absolute values is lower than absolute value of difference, then on each maximum obtain the final inequality.

12 Proof of Theorem 2

Let $f_z(\gamma) = \|V^{\pi^\star_{M,\gamma_{\text{Bw}}}}_{M,\gamma_{\text{Bw}}}(\cdot, z) - V^{\pi^\star_{\widehat{M},\gamma}}_{M,\gamma_{\text{Bw}}}(\cdot, z)\|_\infty$ denote the planning loss given input z. Using Lemma 3, we can bound the difference in planning loss between two different inputs z_i and z_j when planning is conducted with their optimal planning discount factors:

$$\begin{aligned}
f_{z_i}(\gamma^\star(z_i)) - f_{z_j}(\gamma^\star(z_j)) &= f_{z_i}(\gamma^\star(z_i)) - f_{z_i}(\gamma^\star(z_j)) \\
&\quad + f_{z_i}(\gamma^\star(z_j)) - f_{z_j}(\gamma^\star(z_j)) \\
&\leqslant f_{z_i}(\gamma^\star(z_j)) - f_{z_j}(\gamma^\star(z_j)) \\
&\leqslant 2 \max_{\pi:\bar{\mathcal{S}}\mapsto\mathcal{A}} d^\pi_{\text{input},\gamma_{\text{Bw}}}(z_i, z_j)
\end{aligned}$$

and (for the other side):

$$\begin{aligned}
f_{z_j}(\gamma^\star(z_j)) - f_{z_i}(\gamma^\star(z_i)) &= f_{z_j}(\gamma^\star(z_j)) - f_{z_j}(\gamma^\star(z_i)) \\
&\quad + f_{z_j}(\gamma^\star(z_i)) - f_{z_i}(\gamma^\star(z_i)) \\
&\leqslant 2 \max_{\pi:\bar{\mathcal{S}}\mapsto\mathcal{A}} d^\pi_{\text{input},\gamma_{\text{Bw}}}(z_i, z_j),
\end{aligned}$$

which leads to the desired result:

$$|f_{z_i}(\gamma^\star(z_i)) - f_{z_j}(\gamma^\star(z_j))| \leqslant 2 \max_{\pi:\bar{\mathcal{S}}\mapsto\mathcal{A}} d^\pi_{\text{input},\gamma_{\text{Bw}}}(z_i, z_j). \tag{22}$$

For the rest of the proof, we make use of strong convexity. If $f_z(\gamma)$ is μ-strongly convex, we have that for any γ and any $\gamma_0 \in [0, \gamma_{\text{Bw}}]$:

$$f_z(\gamma) \geqslant f_z(\gamma_0) + \langle \nabla f_z(\gamma_0), \gamma - \gamma_0 \rangle + \frac{\mu}{2}|\gamma - \gamma_0|^2. \tag{23}$$

Since $\gamma^\star(z)$ minimizes the planning loss $f_z(\gamma)$, setting $\gamma_0 = \gamma^\star(z)$ leads to a zero derivative $\nabla f_z(\gamma^\star(z)) = 0$ because the planning loss is minimized. We therefore have:

$$f_z(\gamma) \geqslant f_z(\gamma^\star(z)) + \frac{\mu}{2}|\gamma - \gamma^\star(z)|^2. \tag{24}$$

Now by taking $\gamma = \gamma^\star(z_j)$, we can get:

$$f_{z_i}(\gamma^\star(z_j)) \geqslant f_{z_i}(\gamma^\star(z_i)) + \frac{\mu}{2}|\gamma^\star(z_j) - \gamma^\star(z_i)|^2. \tag{25}$$

By rewriting and using triangle inequality along with Equation 22, we obtain Theorem 2:

$$\begin{aligned}
|\gamma^\star(z_i) - \gamma^\star(z_j)| &\leqslant \sqrt{\frac{2}{\mu}|f_{z_i}(\gamma^\star(z_j)) - f_{z_i}(\gamma^\star(z_i))|} \\
&\leqslant \sqrt{\frac{2}{\mu}|(f_{z_i}(\gamma^\star(z_j)) - f_{z_j}(\gamma^\star(z_j))) + (f_{z_j}(\gamma^\star(z_j)) - f_{z_i}(\gamma^\star(z_i)))|} \\
&\leqslant \sqrt{\frac{2}{\mu}|f_{z_i}(\gamma^\star(z_j)) - f_{z_j}(\gamma^\star(z_j))| + |f_{z_j}(\gamma^\star(z_j)) - f_{z_i}(\gamma^\star(z_i))|} \\
&\leqslant \sqrt{\frac{8\max_{\pi:\bar{\mathcal{S}}\mapsto\mathcal{A}} d^{\pi}_{\text{input},\gamma_{\text{Bw}}}(z_i, z_j)}{\mu}}.
\end{aligned}$$

13 MOPO hyperparameters

Hyperparameters play a crucial role in the training of a model-based RL agent. For reproducibility, we therefore list all our hyperparameters here and in the open source code.

Hyperparameter	**Value**
Uncertainty λ	0.5
Proportion of real data for sampling	0.1
Batch Size	128
Starting $\log \alpha$ for SAC	vector of 0
Target Entropy	-1
Target change frequency τ	0.01
Learning rate	10^{-4}
Hold out ratio for dynamics	0.2
Patience for dynamics	50
Hidden size (dynamics)	256
Hidden size (policy)	256
Ensemble Size	3
Max epochs dynamics	500
Steps per epoch	1000
Rollout frequency (steps per rollout)	1000
Rollout length	200
Model retain epochs (for model-buffer)	5
Model rollout batch size	1000
MOPO epochs	50

Table 1: MOPO hyperparameters used in the reward zone CartPole-V1 experiment.