Towards an Understanding of Decision-Time vs. Background Planning in Model-Based Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

In model-based reinforcement learning, an agent can leverage a learned model to improve its way of behaving in different ways. Two of the prevalent approaches are decision-time planning and background planning. In this study, we are interested in *understanding* under what conditions and in which settings one of these two planning styles will perform better than the other. After viewing them in a unified way through the lens of dynamic programming, we first consider the simplest instantiations of these planning styles and provide theoretical results and hypotheses on which one will perform better in the planning & learning and transfer learning settings. We then consider the modern instantiations of them and provide hypotheses on which one will perform better in the considered settings. Lastly, we perform several experiments to illustrate and validate both our theoretical results and hypotheses. Overall, our findings suggest that even though decision-time planning does not perform as well as background planning in its simplest instantiations, the modern instantiations of it can perform on par or better than the modern instantiations of background planning in both the planning & learning and transfer learning settings.

1 Introduction

It has long been argued that, in order for reinforcement learning (RL) agents to adapt to a variety of changing tasks, they should be able to learn a model of their environment, which allows for counterfactual reasoning and fast re-planning (Russell & Norvig, 2002). Although this is a widely-accepted view in the RL community, the question of *how* to leverage a learned model to perform planning in the first place does not have a widely-accepted and clear answer. In model-based RL, the two prevalent planning styles are decision-time and background planning (Sutton & Barto, 2018), where the agent mainly plans in the moment and in parallel to its interaction with the environment, respectively. In the context of RL, even though these two planning styles have been developed with different scenarios and application domains in mind:

- decision-time planning algorithms (Tesauro, 1994; Tesauro & Galperin, 1996; Silver et al., 2017; 2018) for scenarios in which the exact model of the environment is known and for domains that allow for an adequate computational budget at decision time (such as board games like chess and Go);
- background planning algorithms (Sutton, 1990; 1991; Łukasz Kaiser et al., 2020; Hafner et al., 2021) for scenarios in which the exact model is to be learned through pure interaction with the environment and for domains that are agnostic to the response time of the agent (such as gridworld environments and video games),

recently, with the introduction of the ability to learn a model through pure interaction (Schrittwieser et al., 2020), decision-time planning algorithms have been applied to the same scenarios and domains as their background planning counterparts (see e.g., Schrittwieser et al. (2020) and Hamrick et al. (2021) which both evaluate a decision-time planning algorithm, MuZero, on Atari games). However, it still remains unclear under *what* conditions and in *which* settings one of these planning styles will perform better than the other.

In this study, we attempt to provide an answer to one aspect of this question. Specifically, we are interested in answering the following question:

Using the discounted return as the performance measure, under what conditions and in which settings will one planning style perform better than the other?

To answer this question, we first start by abstracting away from the specific implementation details of the two planning styles and view them in a unified way through the lens of dynamic programming. Then, we consider the simplest instantiations of these planning styles and based on their dynamic programming interpretations, provide theoretical results and hypotheses on which one will perform better in the planning & learning and transfer learning settings. We then consider the modern instantiations of these two planning styles and based on both their dynamic programming interpretations and implementation details, provide hypotheses on which one will perform better in the considered settings. Lastly, we perform experiments with both instantiations of these planning styles to illustrate and validate our theoretical results and hypotheses. Overall, our results suggest that even though decision-time planning does not perform as well as background planning in its simplest instantiations, due to (i) the improvements in the way planning is performed, (ii) the usage of only real experience in the updates of the value estimates, and (iii) the ability to improve upon the previously obtained policy at test time, the modern instantiations of it can perform on par or better than their background planning counterparts in both the planning & learning and transfer learning settings. We hope that our findings will help the RL community towards developing a better understanding of how the two planning styles compare against each other and stimulate research in improving them in potentially interesting ways.

2 Background

Reinforcement Learning. In RL (Sutton & Barto, 2018), an agent A interacts with its environment E through a sequence of actions to maximize its long-term cumulative reward. Here, the environment is usually modeled as a Markov decision process $E = (\mathcal{S}_E, \mathcal{A}_E, p_E, r_E, d_E, \gamma)$, where $\mathcal{S}_E$ and $\mathcal{A}_E$ are the (finite) set of states and actions, $p_E : \mathcal{S}_E \times \mathcal{A}_E \times \mathcal{S}_E \to [0, 1]$ is the transition distribution, $r_E : \mathcal{S}_E \times \mathcal{A}_E \times \mathcal{S}_E \to \mathbb{R}$ is the reward function, $d_E : \mathcal{S}_E \to [0, 1]$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. At each time step t, after taking an action $a_t \in \mathcal{A}_E$, the environment's state transitions from $s_t \in \mathcal{S}_E$ to $s_{t+1} \in \mathcal{S}_E$, and the agent receives an observation $o_{t+1} \in \mathcal{O}_E$ and an immediate reward r_t. As there is usually no prior access to the states in $\mathcal{S}_E$ and as the observations in $\mathcal{O}_E$ are usually high-dimensional, the agent has to operate on its own state space $\mathcal{S}_A$, which is generated by its own value encoder $\phi : \mathcal{O}_E \to \mathcal{S}_A$. The goal of the agent is to jointly learn a value encoder ϕ and a value estimator $Q : \mathcal{S}_A \to \mathbb{R}^{|\mathcal{A}_E|}$ that induces a policy $\pi \in \Pi \equiv \{\pi | \pi : \mathcal{S}_A \times \mathcal{A}_E \to [0, 1]\}$, maximizing $E_{\pi, p_E}[\sum_{t=0}^{\infty} \gamma^t r_E(S_t, A_t, S_{t+1}) | S_0 \sim d_E]$. For convenience, we will refer to the composition of ϕ and Q as simply the value estimator.

Model-Free vs. Model-Based RL. The two main ways of achieving this goal are through the use of model-free and model-based RL methods. In model-free RL, there is just a learning phase and the gathered experience is mainly used in improving the value estimator of the agent. In model-based RL however, there are two alternating phases: the learning and planning phases.[1] In the learning phase, in contrast to model-free RL, the gathered experience is mainly used in jointly learning a model encoder $\varphi : \mathcal{O}_E \to \mathcal{S}_M$ and a model $m \in \mathcal{M} \equiv \{(p_M, r_M, d_M) | p_M : \mathcal{S}_M \times \mathcal{A}_E \times \mathcal{S}_M \to [0, 1], r_M : \mathcal{S}_M \times \mathcal{A}_E \times \mathcal{S}_M \to \mathbb{R}, d_M : \mathcal{S}_M \to [0, 1]\}$, where $\mathcal{S}_M$ is the state space of the agent's model, and optionally, as in model-free RL, the experience may also be used in improving the value estimator.[2] For convenience, we will refer to the composition of φ and m as simply the model. In the planning phase, the learned model m is then used for simulating experience, either to be used alongside real experience in improving the value estimator or just to be used in selecting actions at decision time. Note that in general $\varphi \neq \phi$ and thus $\mathcal{S}_M \neq \mathcal{S}_A$. Also note that in the model-based RL literature it is usually implicitly assumed that $\mathcal{S}_E \subseteq \mathcal{S}_M$, which implies that the agent's model is capable of exactly modeling what is going on underneath the environment.

[1] Note that even though some model-based RL algorithms, such as Tesauro & Galperin (1996); Silver et al. (2017; 2018), do not employ a model learning phase and make use of an a priori given exact model, in this study, we will only study versions of them in which the model has to be learned from pure interaction with the environment.

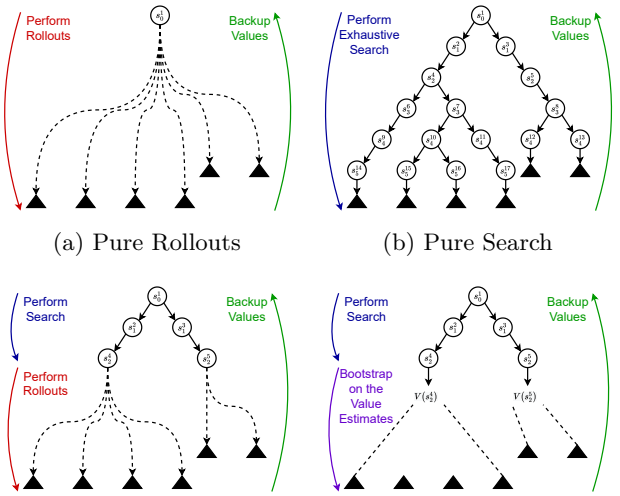
[2] Note that the learned model can both be in a parametric or non-parametric form (see van Hasselt et al. (2019)).

Planning Styles in Model-Based RL. In model-based RL, the two prevalent planning styles are decision-time and background planning (see Ch. 8 of Sutton & Barto (2018)).[3] Decision-time planning is performed as a computation whose output is the selection of a single action for the current state. This is often done by unrolling the model forward from the current state to compute local value estimates, which are then usually discarded after action selection. Here, planning is performed independently for *every* encountered state and it is mainly performed in an *online* fashion, though it may also contain offline components. In contrast, background planning is performed by continually improving a cached value estimator, on the basis of simulated experience from the model, often in a global manner. Action selection is then quickly done by querying the value estimator at the current state. Unlike decision-time planning, background planning is often performed in a purely *offline* fashion, in parallel to the agent-environment interaction, and thus is *not* necessarily focused on the current state: well before action selection for any state, planning plays its part in improving the value estimates in many other states.

For convenience, in this study, we will refer to all model-based RL algorithms that have an online planning component as decision-time planning algorithms (see e.g., Tesauro (1994); Tesauro & Galperin (1996); Silver et al. (2017; 2018); Schrittwieser et al. (2020); Zhao et al. (2021)), and will refer to the rest as background planning algorithms (see e.g., Sutton (1990; 1991); Łukasz Kaiser et al. (2020); Hafner et al. (2021); Zhao et al. (2021)). Note that, regardless of the style, any type of planning can be viewed as a procedure $f : (\mathcal{M}, \Pi) \rightarrow \Pi$ that takes a model m and a policy π^i as input and returns an improved policy π^o_m, according to m, as output.

Categorization within the Two Planning Styles. Starting with decision-time planning, depending on how much search is performed, decision-time planning algorithms can be studied under three main categories:

1. Decision-time planning algorithms that perform no search (see e.g., Tesauro & Galperin (1996) and Alg. 1)
2. Decision-time planning algorithms that perform pure search (see e.g., Campbell et al. (2002) and Alg. 2)
3. Decision-time planning algorithms that perform some amount of search (see e.g., Silver et al. (2017; 2018); Schrittwieser et al. (2020) and Alg. 5)



(a) Pure Rollouts (b) Pure Search

(c) Search + Rollouts (d) Search + Bootstrapping

Figure 1: The different planning styles within decision-time planning in which planning is performed (a) by purely performing rollouts, (b) by purely performing search, (c) by performing rollouts after performing some amount of search and (d) by bootstrapping on the value estimates after performing some amount of search. The subscripts and superscripts on the states indicate the time steps and state identifiers, respectively. The black triangles indicate the terminal states.

In the first two of these categories, planning is performed by just running pure rollouts with a fixed or improving policy (see Fig. 1a), and by purely performing search (see Fig. 1b), respectively. In the last one, planning is performed by first performing some amount of search and then either by running rollouts with a fixed or improving policy, by bootstrapping on the cached value estimates of a fixed or improving policy, or by doing both (see Fig. 1c & 1d). Note that while the simple instantiations of decision-time planning fall within the first two categories, the modern instantiations of it fall within the last one. Also note that, while planning is performed with only a single parametric model in the first two categories, it is usually performed with both a parametric and non-parametric (usually a replay buffer, see van Hasselt et al. (2019))

[3]Although some new planning styles have been proposed in the transfer learning literature (see e.g., Barreto et al. (2017; 2019; 2020); Alver & Precup (2022)), these approaches can also be viewed as performing some form of decision-time planning with pre-learned models.

model in the last one. We refer the reader to Bertsekas (2021) for more details on the different categories of decision-time planning algorithms.

Moving on to background planning, as all background planning algorithms (see e.g., Sutton (1990; 1991); Łukasz Kaiser et al. (2020); Hafner et al. (2021) and Alg. 3, 4, 6) perform planning by continually improving a cached value estimator throughout the model learning process, we do not study them under different categories. However, we again note that while the simplest instantiations of background planning perform planning with a single parametric model (see e.g., Alg. 3 & 4), the modern instantiations of it perform planning with both a parametric and non-parametric (again usually a replay buffer) model (see e.g., Alg. 6).

3 A Unified Dynamic Programming View of the Two Planning Styles

In this section, we abstract away from the specific implementation details, such as whether policy improvement is done locally or globally, or whether planning is performed in an online or offline manner, and view the two planning styles in a unified way through the lens of dynamic programming (Bertsekas & Tsitsiklis, 1996). More specifically, we view decision-time and background planning through the lens of the well-known policy iteration (PI) algorithm.[4] In this framework, the two planning styles can be viewed as follows:

- Decision-time planning algorithms that perform no search can be considered as performing one-step PI on top of a fixed or improving policy at every time step as at each time step they compute a π^o_m by first running many rollouts in m with a fixed or improving π^i to evaluate the current state (which corresponds to policy evaluation), and then by selecting the most promising action (which corresponds to policy improvement). Similarly, decision-time planning algorithms that perform pure search can be considered as performing PI until convergence (which we call full PI) at every time step as at each time step they disregard π^i and compute a π^o_m by first performing exhaustive search in m to obtain the optimal values at the current state, and then by selecting the most promising action. Finally, decision-time planning algorithms that perform some amount of search can be considered as performing an amount of PI that is between one-step and full PI on top of a fixed or improving policy at every time step as they are at the intersection of decision-time planning algorithms that perform no search and pure search.

- All background planning algorithms can be considered as performing an amount of PI that is usually less than one-step PI on top of an improving policy at every time step as at each time step they compute a π^o_m by gradually improving π^i on the basis of simulated experience from m. However, if the learned model m converges in the model learning process, background planning algorithms can be considered as performing an amount of PI that is eventually equivalent to full PI as the continual improvements to π^i at each time step would eventually lead to an improvement that is equivalent to performing full PI.

Note that while the PI view of decision-time planning abstracts a planning process that focuses on the agent's current state, the PI view of background planning abstracts a one that is dispersed across the agent-environment interaction. Also note that, in Sec. 2, we have pointed out that some decision-time and background planning algorithms perform planning with both a parametric and non-parametric model, which can make it hard for them to be viewed through the proposed unified framework. However, if one considers the two separate models as a single combined model, then these algorithms can also be viewed straightforwardly in our proposed framework. We refer the reader to App. A for a broader discussion.

4 Decision-Time vs. Background Planning

In this study, we are interested in understanding under what conditions and in which settings will one planning style perform better than the other. Thus, we start by formally defining a performance measure that

[4]We choose policy iteration over value iteration as it better describes how planning is performed in decision-time planning (see Bertsekas (2021)), and it is also useful in describing the planning process in background planning.

will be used in comparing the two planning styles of interest. Given an arbitrary model $m = (p, r, d) \in \mathcal{M}$, let us define the performance of an arbitrary policy $\pi \in \Pi$ in it as follows:

$$J_m^\pi \equiv E_{\pi,p}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t, S_{t+1}) \middle| S_0 \sim d\right]. \tag{1}$$

Note that J_m^π corresponds to the expected *discounted* return of a policy π in model m. Next, we start considering the conditions under which the comparisons will be made: we are interested in both simple scenarios in which the value estimators and models are represented as a table, and in complex ones in which at least the value estimators or models are represented using function approximation.

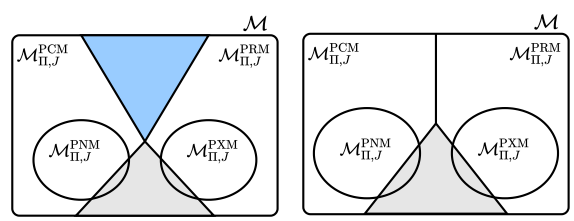
4.1 Partitioning of the Model Space

Before moving on to the comparison of the two planning styles across different settings, we first present a way to partition the space of agent models $\mathcal{M}$ such that it would be possible to understand when will one planning style be guaranteed to perform on par or better than the other. Let us start by defining m^* to be the exact model of the environment. Note that $m^* \in \mathcal{M}$ as $\mathcal{S}_E \subseteq \mathcal{S}_M$ (see Sec. 2). Then, given a policy set $\Pi \subseteq \Pi$, containing at least two policies, and a performance measure J, defined as in (1), depending on the relative performances of the policies in it and in m^*, a model $m \in \mathcal{M}$ can belong to one of the following main classes:

Definition 1 (PCM)**.** *Given a $\Pi \subseteq \Pi$ and a J, let $\mathcal{M}_{\Pi,J}^{\text{PCM}} \equiv \{m \in \mathcal{M} \mid J_{m^*}^{\pi^i} \lesseqgtr J_{m^*}^{\pi^j}$ for all $\pi^i, \pi^j \in \Pi$ satisfying $J_m^{\pi^i} \gtreqless J_m^{\pi^j}\}$. We say that each $m \in \mathcal{M}_{\Pi,J}^{\text{PCM}}$ is a* performance-contrasting model (PCM) *of m^* with respect to Π and J.*

Definition 2 (PRM)**.** *Given a $\Pi \subseteq \Pi$ and a J, let $\mathcal{M}_{\Pi,J}^{\text{PRM}} \equiv \{m \in \mathcal{M} \mid J_{m^*}^{\pi^i} \gtreqless J_{m^*}^{\pi^j}$ for all $\pi^i, \pi^j \in \Pi$ satisfying $J_m^{\pi^i} \gtreqless J_m^{\pi^j}\}$. We say that each $m \in \mathcal{M}_{\Pi,J}^{\text{PRM}}$ is a* performance-resembling model (PRM) *of m^* with respect to Π and J.*

Informally, given any two policies in Π and a J, a model m is a PCM of m^* iff the policy that performs on par or better in it performs on par or worse in m^*, and it is a PRM of m^* iff the policy that performs on par or better in it also performs on par or better in m^*. Note that m can both be a PCM and a PRM of m^* iff the two policies perform on par in both m and m^*. If Π contains at least one of the optimal policies for m, then m can also belong to one of the following specialized classes:



(a) General (b) Of interest

Figure 2: (a) The general partitioning and (b) the partitioning of interest of $\mathcal{M}$, for a given Π and J. The gray and blue regions indicate $\mathcal{M}_{\Pi,J}^{\text{PCM}} \cap \mathcal{M}_{\Pi,J}^{\text{PRM}}$ and $\mathcal{M} \setminus (\mathcal{M}_{\Pi,J}^{\text{PCM}} \cup \mathcal{M}_{\Pi,J}^{\text{PRM}})$, respectively.

Definition 3 (PNM)**.** *Given a $\Pi \subseteq \Pi$ and a J, let $\mathcal{M}_{\Pi,J}^{\text{PNM}} \equiv \{m \in \mathcal{M}_{\Pi,J}^{\text{PCM}} \mid J_{m^*}^{\pi_m^*} = \min_{\pi \in \Pi} J_{m^*}^{\pi}$ for all $\pi_m^* \in \Pi\}$, where π_m^* denotes the optimal policies in m. We say that each $m \in \mathcal{M}_{\Pi,J}^{\text{PNM}}$ is a* performance-minimizing model (PNM) *of m^* with respect to Π and J.*

Definition 4 (PXM)**.** *Given a $\Pi \subseteq \Pi$ and a J, let $\mathcal{M}_{\Pi,J}^{\text{PXM}} = \{m \in \mathcal{M}_{\Pi,J}^{\text{PRM}} \mid J_{m^*}^{\pi_m^*} = \max_{\pi \in \Pi} J_{m^*}^{\pi}$ for all $\pi_m^* \in \Pi\}$, where π_m^* denotes the optimal policies in m. We say that each $m \in \mathcal{M}_{\Pi,J}^{\text{PXM}}$ is a* performance-maximizing model (PXM) *of m^* with respect to Π and J.*

Informally, given a subset of Π that contains the optimal policies for a model m, and a J, m is a PNM of m^* iff all of the optimal policies result in the worst possible performance in m^*, and it is a PXM of m^* iff all them result in the best possible performance in m^*. Note that PNMs are a subclass of PCMs and PXMs are a subclass of PRMs. Also note that all definitions above are agnostic to how the models are represented, i.e., whether if they are represented through tables or through the use of function approximation.

Fig. 2a illustrates how $\mathcal{M}$ is generally partitioned for an arbitrary Π and J. Note that given a fixed J, the relative sizes of the model classes solely depend on Π. For instance, as Π gets larger, the relative sizes of $\mathcal{M}_{\Pi,J}^{\text{PCM}}$ and $\mathcal{M}_{\Pi,J}^{\text{PRM}}$ shrink, because with every policy that is added to Π, the number of criteria that a model

must satisfy to be a PCM or PRM increases, which reduces the odds of an arbitrary model in $\mathcal{M}$ being in $\mathcal{M}^{\text{PCM}}_{\Pi,J}$ or $\mathcal{M}^{\text{PRM}}_{\Pi,J}$ (see the blue region in Fig. 2a). And, as Π gets smaller, the relative sizes of $\mathcal{M}^{\text{PCM}}_{\Pi,J}$ and $\mathcal{M}^{\text{PRM}}_{\Pi,J}$ grow, and eventually fill up the entire space, when Π contains only two policies. Fig. 2b illustrates the partitioning in this scenario. Since we are only interested in comparing the policies of two planning styles, the Π of interest has a size of two, i.e. $|\Pi| = 2$, and thus we have a partitioning as in Fig. 2b which covers the *whole* model space $\mathcal{M}$.

Illustrative Example. As an illustration of the model classes defined above, let us start by considering the Simple Gridworld environment depicted in Fig. 3, in which the agent spawns in state S and has to navigate to the goal state depicted by G. At each time step, the agent receives an (x, y) pair, indicating its position, and based on this, selects an action that moves it to one of the four neighboring cells with a slip probability of 0.05. The agent receives a negative reward that is linearly proportional to its distance from G and a reward of +10 if it reaches G. In this environment, given a policy set Π that contains the policies of the two planning styles and the performance measure J, examples of PCMs and PRMs can be tabular models with goal states of $\{\text{G}_n\}_{n=2}^{5}$ and $\{\text{G}_n\}_{n=6}^{9}$, respectively. And, assuming that Π contains at least one of the optimal policies, examples of a PNM and a PXM can be tabular models with goal states of G_1 and G, respectively.

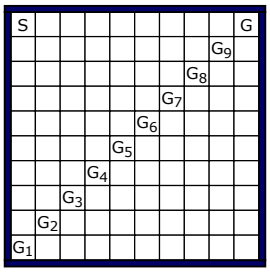


Figure 3: The Simple Gridworld environment.

Even though we have provided definitions for four different model classes, throughout the rest of this study, we will only provide theoretical results and hypotheses for the scenarios in which the models of the two planning styles converge to PRMs and PXMs as the model learning process pushes their models to become either PRMs or PXMs, even if they start as PCMs or PNMs. We have only provided definitions for PCMs and PNMs to paint a complete picture of the space of possible models.

4.2 Simplest Instantiations of the Two Planning Styles

We are now ready to discuss when will one planning style perform better than the other across different settings. For easy analysis, we start by considering the simplest instantiations of the two planning styles, which can be found in Ch. 8 of Sutton & Barto (2018) and in which both the value estimators and models are represented as a table. More specifically, for decision-time planning we study a version of the online Monte-Carlo planning algorithm of Tesauro & Galperin (1996) in which a parametric model is learned from experience (see Alg. 1) and for background planning we study a simplified version of the Dyna-Q algorithm of Sutton (1990; 1991) in which the value estimator is updated by using samples from only the model and not the environment (see Alg. 4). Note that for the simplest instantiation of decision-time planning, we choose to study an algorithm that performs no search, and not a one that performs pure search (see Sec. 2), as the latter ones require a significant amount of computational budget at decision-time and thus are not practically applicable to most scenarios. We refer the reader to App. B for a discussion on why we consider these specific versions.

In the unified framework that we introduced in Sec. 3, these algorithms can be viewed as follows:

- As the decision-time planning algorithm performs planning by first running many rollouts with a fixed policy in the model and then by selecting the most promising action, it can be considered as performing one-step PI on top of a fixed policy at every time step.
- And, as the background planning algorithm performs planning by continually improving a value estimator at every time step with samples from the model, it can be viewed as performing an amount of PI that is eventually equivalent full PI when its learned model converges.

Note that, although we only consider these specific instantiations, as long as decision-time planning corresponds to taking a smaller or on par policy improvement step than background planning (which is the case in simple instantiations that are practically applicable to most scenarios), the formal results and hypotheses that we provide in this section would hold regardless of the choice of instantiation.

Before considering different settings, let us define the following policies that will be useful in referring to the input and output policies of the two planning styles:

Definition 5 (Base, Rollout (Bertsekas, 2021) and Certainty-Equivalence (Jiang et al., 2015) Policies)**.** *The* base policy $\pi^b \in \Pi$ *is the policy that is used in initiating PI. Given a base policy* π^b *and a model* $m \in \mathcal{M}$*, the* rollout policy $\pi^r_m \in \Pi$ *is the policy obtained after performing one-step of PI on top of* π^b *in* m*, and the* certainty-equivalence policy $\pi^{ce}_m \in \Pi$ *is the policy obtained after performing full PI in* m*.*

In the rest of this section, we will refer to the policies generated by the simplest instantiations of decision-time and background planning with model m as π^r_m and π^{ce}_m, respectively.

4.2.1 Planning & Learning Setting

Fairest Scenario. For the fairest possible comparison, we start by considering the scenario in which the two planning styles would perform planning with the same model $m \in \mathcal{M}$ that is to be learned in the model learning process. In this setting, we can prove the following statement:

Proposition 1. *Let* $m \in \mathcal{M}$ *be a PRM of* m^* *with respect to* $\Pi = \{\pi^r_m, \pi^{ce}_m\} \subseteq \Pi$ *and* J*. Then,* $J^{\pi^{ce}_m}_{m^*} \geq J^{\pi^r_m}_{m^*}$*.*

Due to space constraints, we defer the proofs to App. C. Prop. 1 implies that, given $\Pi = \{\pi^r_m, \pi^{ce}_m\}$ and J, decision-time planning will perform on par or worse than background planning if m converges to a PRM. Note that even though this result would not be guaranteed to hold if function approximation was introduced in the value estimator representations (as in this case, there would be no guarantee that full PI will result in a better policy than one-step PI in m, which is a basic result from dynamic programming (Bertsekas & Tsitsiklis, 1996)), if one were to use approximators with good generalization capabilities, i.e., approximators that assign the same value to similar observations, we would expect a similar performance trend to hold.

Common Scenario. In common comparison scenarios, instead of restricting the two planning styles to perform planning with the same model, the comparison is usually done by allowing the two planning styles to perform planning with their own models that are again to be learned in the model learning process. In this scenario, as different trajectories are likely to be used in the model learning process, the encountered models of the two planning styles, which we denote as $\bar{m} \in \mathcal{M}$ and $\bar{\bar{m}} \in \mathcal{M}$ for decision-time and background planning, respectively, are also likely to be different. Thus, even though coming up with a theoretical result that is as strong as Prop. 1 is not possible, we can still prove the following statement:

Proposition 2. *Let* $\bar{m} \in \mathcal{M}$ *be any model of* m^* *with respect to* $\bar{\Pi} = \{\pi^r_{\bar{m}}, \pi^{ce}_{\bar{m}}\} \subseteq \Pi$ *and* J*, and let* $\bar{\bar{m}} \in \mathcal{M}$ *be a PXM of* m^* *with respect to* $\bar{\bar{\Pi}} = \{\pi^r_{\bar{\bar{m}}}, \pi^{ce}_{\bar{\bar{m}}}\} \subseteq \Pi$ *and* J*. Then,* $J^{\pi^{ce}_{\bar{\bar{m}}}}_{m^*} \geq J^{\pi^r_{\bar{m}}}_{m^*}$*.*

Prop. 2 implies that, given $\bar{\Pi} = \{\pi^r_{\bar{m}}, \pi^{ce}_{\bar{m}}\}$, $\bar{\bar{\Pi}} = \{\pi^r_{\bar{\bar{m}}}, \pi^{ce}_{\bar{\bar{m}}}\}$ and J, decision-time planning will perform on par or worse than background planning if $\bar{\bar{m}}$ converges to a PXM. Also note again that even though Prop. 2 would not be guaranteed to hold if function approximation was used in the value estimator representations (due to the reason discussed in the fairest comparison scenario), if one were to use approximators with good generalization capabilities, we would again expect a similar performance trend to hold.

4.2.2 Transfer Learning Setting

Adaptation Scenario. Although there are many different settings in transfer learning (Taylor & Stone, 2009), for easy analysis, we start by considering a simple and commonly used one for tabular agents in which there is only one training task and one subsequent test task, differing only in their reward functions, and in which the agent's transfer ability is measured by how fast it adapts to the test task after being trained on the training task (more challenging settings will be considered in the next section). We refer to this transfer setting as the adaptation setting. In this setting, we would expect the results of the planning & learning setting to hold directly, as instead of a single one, there are now two consecutive planning & learning settings.

4.3 Modern Instantiations of the Two Planning Styles

We now consider the modern instantiations of the two planning styles in which both the value estimators and models are represented with neural networks. More specifically, we study the decision-time and background

planning algorithms in Zhao et al. (2021) (see Alg. 5 & 6) as they are *generic* algorithms that are reflective of many of the properties of their state-of-the-art counterparts (see e.g., MuZero (Schrittwieser et al., 2020), SimPLe (Łukasz Kaiser et al., 2020), DreamerV2 (Hafner et al., 2021)) and their code for both planning styles is publicly available. We refer the reader to App. D for a broader discussion on why we choose these algorithms.

In the unified framework that we introduced in Sec. 3, these algorithms can be viewed as follows:

- As the decision-time planning algorithm performs planning by first performing some amount of search with a parametric model and then by bootstrapping on the value estimates of a continually improving policy, it can be considered as performing more than one-step but less than full PI on top of an improving policy with a combined model $\bar{m}_c$ at every time step.
- And, as the background planning algorithm performs planning by continually improving a value estimator at every time step with samples from both a parametric and non-parametric (a replay buffer) model, it can be viewed as performing an amount of PI that is eventually equivalent to full PI with a combined model $\bar{\bar{m}}_c$, when $\bar{\bar{m}}_c$ converges.

Note that although we only consider these specific instantiations, the formal results and hypotheses we provide in this section are also generally applicable to most state-of-the-art model-based RL algorithms (see e.g., Schrittwieser et al., 2020; Łukasz Kaiser et al., 2020; Hafner et al., 2021), as the algorithms in Zhao et al. (2021) are reflective of many of their properties (see App. D).

4.3.1 Planning & Learning Setting

Simplified Scenario. To ease the analysis, let us start by considering a simplified scenario in which both the value estimators and models of the modern instantiations of the two planning styles are represented as a table. Let us also define the *improved rollout policy* to be as follows:

Definition 6 (Improved Rollout Policy)**.** *Given a base policy π^b and a model $m \in \mathcal{M}$, the* improved rollout policy $\pi^{r+}_m \in \Pi$ *is the policy obtained after performing more than one-step but less than full PI on top of π^b in m.*

And, let us also refer to the policies generated by the modern instantiations of decision-time and background planning with models $\bar{m}_c$ and $\bar{\bar{m}}_c$ as $\pi^{r+}_{\bar{m}_c}$ and $\pi^{ce}_{\bar{\bar{m}}_c}$, respectively. Then, using $\pi^{r+}_{\bar{m}_c}$ and $\pi^{ce}_{\bar{\bar{m}}_c}$ in place of $\pi^{r}_{\bar{m}}$ and $\pi^{ce}_{\bar{\bar{m}}}$, respectively, we would expect Prop. 2 to hold exactly as decision-time planning still corresponds taking a smaller or on par policy improvement step than background planning. However, as decision-time planning now corresponds to performing more than one-step PI, we would expect the performance gap between the two planning styles to reduce in their modern instantiations. Moreover, we would expect this gap to gradually close if both $\bar{m}_c$ and $\bar{\bar{m}}_c$ converge to PXMs, as the use of an improving base policy for decision-time planning would result in a continually improving performance that gets closer to the one of background planning.

Original Scenario. Coming back to the original scenario in which both the value estimators and models of the two planning styles are represented with neural networks, we would expect a similar performance trend to hold as neural networks are approximators with good generalization capabilities. However, note that this expectation is solely based on our dynamic programming view of the two planning styles and thus does not take into consideration the issues that may arise due to the implementation details of them, which can also play an important role on how they will compare against each other under the scenario of both $\bar{m}_c$ and $\bar{\bar{m}}_c$ converging to PXMs.

In their modern instantiations, both planning styles perform planning by unrolling neural network based parametric models which are known to easily lead to compounding model errors (Talvitie, 2014). Thus, performing effective planning becomes quite difficult for both planning styles. Even though this problem can significantly be mitigated for both of them by unrolling the models for only a few time steps and then bootstrapping on the value estimators for the rest, background planning can also suffer from updating its value estimator with the potentially harmful simulated experience that is generated by its neural network based parametric model (see line 20 in Alg. 6), which can slowdown or even prevent it in reaching optimal (or

close-to-optimal) performance (see e.g., van Hasselt et al. (2019); Jafferjee et al. (2020)). Note that this is not a problem in decision-time planning as it performs updates to its value estimator with only real experience. Thus, we hypothesize that compared to decision-time planning, it is likely for background planning to suffer more in reaching optimal (or close-to-optimal) performance in their modern instantiations.

4.3.2 Transfer Learning Setting

Adaptation and Zero-shot Scenarios. We now consider two commonly used transfer learning settings that are both more challenging than the setting considered in Sec. 4.2. In these settings, there is a distribution of training and test tasks, differing only in their observations. In the first one, the agent's transfer ability is measured by how fast it adapts to the test tasks after being trained on the training tasks, which we again refer to as the adaptation setting (see e.g., Van Seijen et al. (2020)), and in the second one, this ability is measured by the agent's instantaneous performance on the test tasks as it gets trained on the training tasks, which we refer to as the zero-shot setting (see e.g., Zhao et al. (2021); Anand et al. (2022)).

In both settings we would again expect background planning to suffer more in reaching optimal (or close-to-optimal) performance on the training tasks because of the same reasons discussed in the planning & learning setting. Additionally, in the adaptation setting, after the tasks switch from the training tasks to the test tasks, we would expect background planning to suffer more in the adaptation process, as its parametric model, learned on the training tasks, would keep generating experience that resembles the training tasks until it adapts to the test tasks, which in the meantime can lead to harmful updates to its value estimator. Also, in the zero-shot setting, if the learned parametric model of decision-time planning becomes capable of simulating at least a few time steps of the test tasks, we would expect decision-time planning to perform better on the test tasks, as at test time it would be able to improve upon the policy obtained during training by performing online planning. Note that, this is not possible for background planning, as it performs planning in an offline fashion and thus requires additional interaction with the test tasks to improve upon the policy obtained during training, which is not possible in the zero-shot setting.

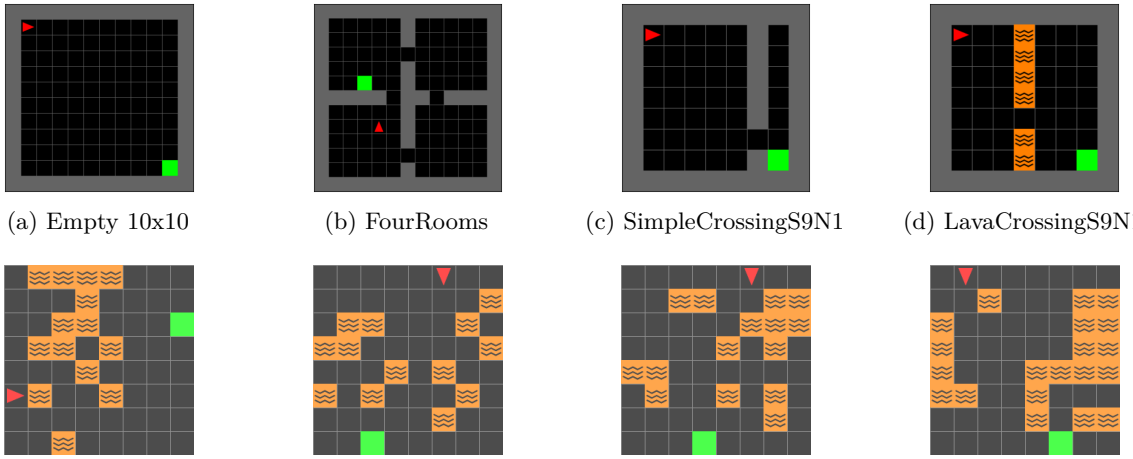
5 Experiments and Results

We now perform experiments to illustrate and validate the formal statements and hypotheses presented in Sec. 4. The experimental details and more detailed results can be found in App. E & F, respectively.

Environmental Details. As a testbed, we make use of the Simple Gridworld environment (see Fig. 3) and of several MinGrid environments (Chevalier-Boisvert et al., 2018) (see Fig. 4). We choose these environments as the optimal policies in them are easy to learn and they allow for designing controlled experiments that are helpful in answering the questions of interest to this study. The details of the Simple Gridworld environment are already presented in Sec. 4.1. In the MiniGrid environments, the agent (depicted in red) has to navigate to the green goal cell, while avoiding the orange lava cells (if there are any). At each time step, the agent receives a grid-based observation that contains its own position and the positions of the goal, wall and lava cells, and based on this, selects an action that either turns it left or right, or steps it forward. If the agent steps on a lava cell, the episode terminates with no reward, and if it reaches the goal cell, the episode terminates with a reward of $+1$. More details on these environments can be found in App. E. Note that while $\mathcal{O}_E = \mathcal{S}_E$ in the Simple Gridworld environment, $\mathcal{O}_E \neq \mathcal{S}_E$ in the MiniGrid environments.

5.1 Experiments with Simplest Instantiations

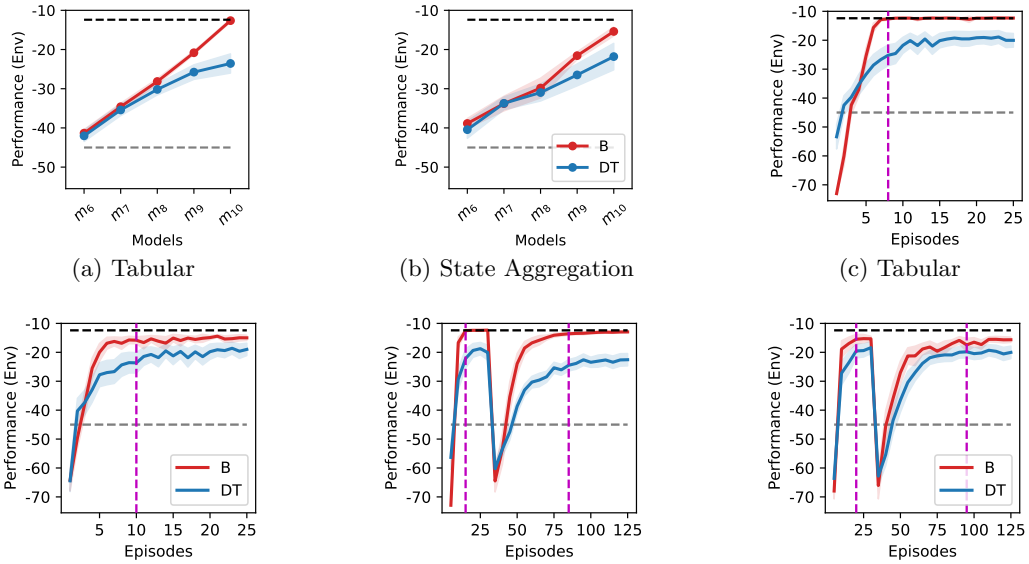
In this section, we perform experiments with the simplest instantiations of decision-time and background planning (see Alg. 1 & 4) on the Simple Gridworld environment to illustrate our theoretical results and hypotheses presented in Sec. 4.2. In addition to the scenario in which both the value estimators and models are represented as tables (where $\phi = \varphi$ are both identity functions, implying $\mathcal{S}_A = \mathcal{S}_M = \mathcal{O}_E$), we also consider the one in which only the model is represented as a table (where only φ is an identity function, only implying $\mathcal{S}_M = \mathcal{O}_E$). In the latter scenario, we use state aggregation in the value estimator representation, i.e., ϕ is a state aggregator. More on the implementation details of these instantiations can be found in App. E.1.



(a) Empty 10x10 (b) FourRooms (c) SimpleCrossingS9N1 (d) LavaCrossingS9N1

(e) RDS Train (Diff: 0.35) (f) RDS Test (Diff: 0.25) (g) RDS Test (Diff: 0.35) (h) RDS Test (Diff: 0.45)

Figure 4: (a-d) The Empty 10x10, FourRooms, SimpleCrossingS9N1 and LavaCrossingS9N1 environments in MiniGrid. (e-h) The training task of difficulty 0.35 and test tasks of difficulties 0.25, 0.35 and 0.45 in the RDS environment (Zhao et al., 2021). Note that the difficulty parameter here controls the density of the lava cells between the agent and the goal cell, and that the test tasks are just transposed versions of the training tasks. Also note that with every reset of the episode, a new lava cell pattern is procedurally generated for both the training and test tasks. More on the details of the RDS environment can be found in Zhao et al. (2021).



(a) Tabular (b) State Aggregation (c) Tabular

(d) State Aggregation (e) Tabular (f) State Aggregation

Figure 5: The performance of the simplest instantiations of decision-time (DT) and background (B) planning on the Simple Gridworld environment, in the (a, b, c, d) planning & learning and (e, f) transfer learning settings with tabular and state aggregation value estimator representations. Black & gray dashed lines indicate the performance of the optimal & random policies, respectively. The magenta dashed line in (c, d, e, f) indicates the point after which background planning's model becomes and remains as a PXM. Shaded regions are one standard error over 250 runs.

5.1.1 Planning & Learning Experiments

Fairest Scenario. According to Prop. 1, decision-time planning is guaranteed to perform on par or worse than background planning if their model m converges to a PRM. For empirical verification, we designed a controlled setting in which we trained m to sequentially converge to as set of PRMs: m first converges to the PRMs $\{m_j\}_{j=6}^{9}$ with goal states $\{\mathrm{G}_n\}_{n=6}^{9}$ and then converges to the PXM m_{10} with goal state G (which is also a PRM, see Fig. 3 and App. E.1 for more details on these models). After planning was performed with each of these models, we evaluated the resulting output policies in the environment. Results are shown in Fig. 5a. We can indeed see that decision-time planning performs worse when m converges to a PRM. To see if similar results would hold with approximators that have good generalization capabilities, we also performed the same experiment with state aggregation used in the value estimator representation. Results in Fig. 5b show that a similar trend holds in this case as well.

Common Scenario. Prop. 2 implies that, decision-time planning is guaranteed to perform on par or worse than background planning if the model of background planning converges to a PXM. For empirical verification, we initialized the tabular models of both planning styles as randomized models and let them be updated through interaction to become PXMs. After every episode, we evaluated the resulting output policies in the environment. Results in Fig. 5c show that, as expected, decision-time planning performs worse when the model of background planning converges to a PXM. Again, to see if similar results would hold with approximators that have good generalization capabilities, we also performed experiments with state aggregation used in the value estimator representation. Results in Fig. 5d show that a similar trend holds in this case as well.

5.1.2 Transfer Learning Experiments

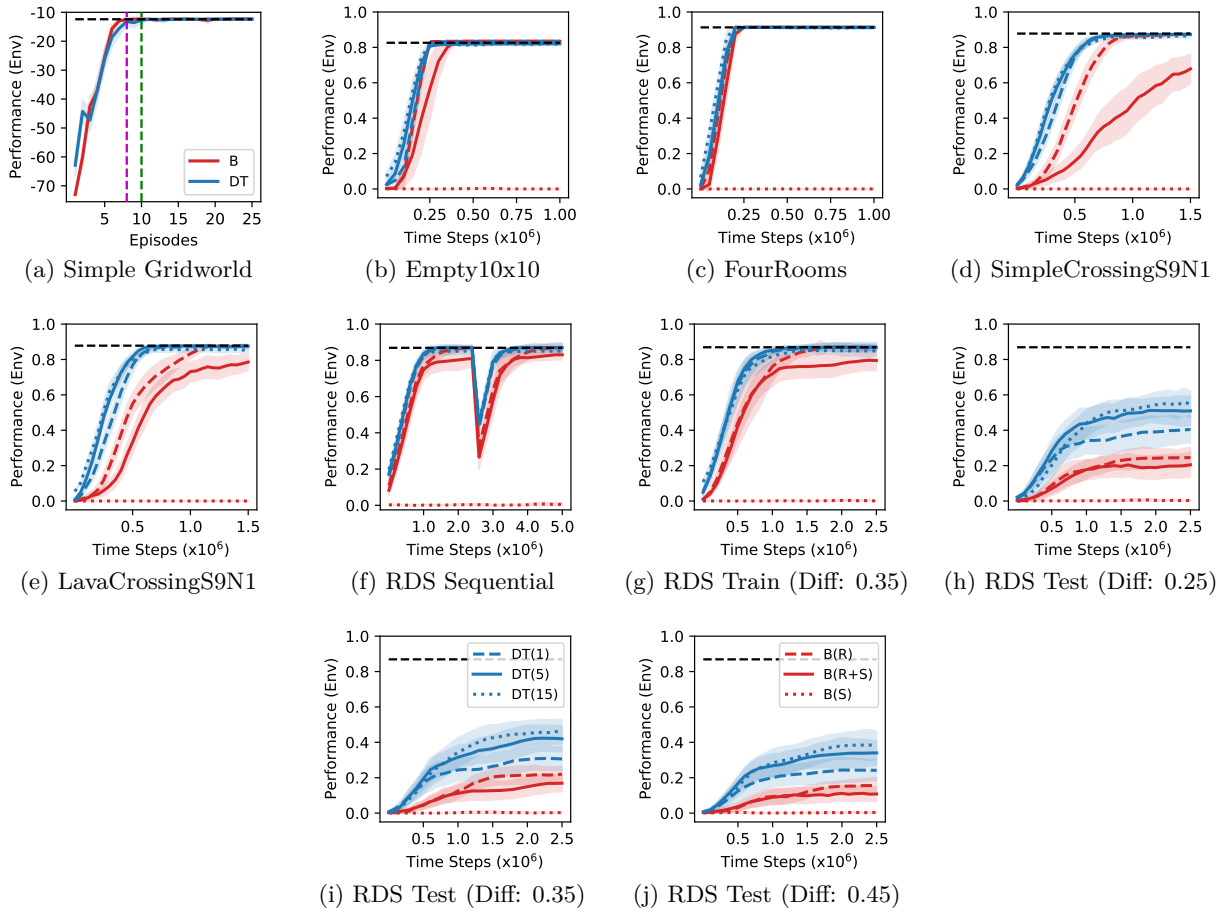
Adaptation Scenario. In Sec. 4.2.2, we argued that the results of the planning & learning setting would hold directly in the considered adaptation setting. For empirical verification, we performed a similar experiment to the one in the planning & learning setting, in which we initialized the tabular models of both planning styles as randomized models and let them be updated to become PXMs. However, differently, after 25 episodes, we now added a subsequent test task with goal state G_1 to the training task (see App. E.1 for the details). In Fig. 5e, we can see that, similar to the planning & learning setting, before the task changes, decision-time planning performs worse when the model of background planning converges to a PXM, and the same happens after the task changes. Results in Fig. 5f show that a similar trend also holds when state aggregation is used in the value estimator representation.

5.2 Experiments with Modern Instantiations

We now perform experiments with the modern instantiations of decision-time and background planning to empirically validate our hypotheses in Sec. 4.3. For the experiments with the Simple Gridworld environment, we consider the former scenario in Sec. 5.1, and for the experiments with MiniGrid environments, we consider the scenario in which both the value estimators and models are represented with neural networks, and in which they share the same encoder (where $\phi = \varphi$ are both learnable parametric functions, implying $\mathcal{S}_A = \mathcal{S}_M \neq \mathcal{O}_E$). More on the implementation details of these instantiations can be found in App. E.2.

5.2.1 Planning & Learning Experiments

Simplified Scenario. In Sec. 4.3.1, we argued that if one were to use tabular value estimators and models in the modern instantiations, the results of the planning & learning section of Sec. 4.2 would hold exactly. Additionally, we also argued that the performance gap between the two planning styles would reduce in their modern instantiations and even gradually close if the combined models of both planning styles converge to PXMs. In order to test these, we implemented simplified tabular versions of the modern instantiations of the two planning styles (see Alg. 7 & 8) and compared them on the Simple Gridworld environment. Results are shown in Fig. 6a. As expected, the results of the planning & learning section of Sec. 4.2 holds exactly, and the performance gap between the two planning styles indeed reduces and gradually closes after the models converge to PXMs.



(a) Simple Gridworld (b) Empty10x10 (c) FourRooms (d) SimpleCrossingS9N1

(e) LavaCrossingS9N1 (f) RDS Sequential (g) RDS Train (Diff: 0.35) (h) RDS Test (Diff: 0.25)

(i) RDS Test (Diff: 0.35) (j) RDS Test (Diff: 0.45)

Figure 6: The performance of the modern instantiations of decision-time (DT) and background (B) planning in the (a-e) planning & learning and (f-j) transfer learning settings with (a) tabular and (b-j) neural network value estimator representations. The black dashed lines indicate the performance of the optimal policy in the corresponding environment. The green and magenta dashed line in (a) inidicates the point after which decision-time and background planning's models become and remain as PXMs, respectively. Shaded regions are one standard error over (a) 250 and (b-j) 100 runs.

Original Scenario. We then argued that although the usage of neural networks in the representation of the value estimators and models of the two planning styles would not break the performance trend, their implementation details can also play an important role on how they would compare against each other. We hypothesized that although both planning styles would suffer from compounding model errors, background planning would additionally suffer from updating its value estimator with the potentially harmful simulated experience, and thus it is likely to suffer more in reaching optimal (or close-to-optimal) performance.

To test these hypotheses, we compared the modern instantiations of the two planning styles (see Alg. 5 & 6) on four MiniGrid environments: Empty 10x10, FourRooms, SimpleCrossingS9N1 and LavaCrossingS9N1. In order to test the effect of compounding model errors in decision-time planning, we performed experiments in which the parametric model is unrolled for 1, 5 and 15 time steps during search, denoted as DT(1), DT(5) and DT(15). Note that compounding model errors are not a problem for the considered background planning algorithm, as its parametric model is only unrolled for a single time step. Also, in order to test the effect of updating the value estimator with simulated experience in background planning, we performed experiments in which its value estimator is updated with only real, both real and simulated, and only simulated experience, denoted as B(R), B(R+S) and B(S). See App. E.2 for the details of these experiments. The results, in Fig. 6b-6e, show that even though the mean performance of decision-time planning degrades

slightly with the increase in compounding model errors, this can indeed be easily mitigated by decreasing the search budget. However, as can be seen, even without any compounding model errors, just the usage of simulated experience alone can indeed slowdown, or even prevent (as in B(S)), background planning in reaching optimal (or close-to-optimal) performance, especially in relatively hard-to-model environments such as SimpleCrossingS9N1 and LavaCrossingS9N1.

5.2.2 Transfer Learning Experiments

In Sec. 4.3.2, we first argued that background planning would again suffer more in reaching optimal (or close-to-optimal) performance on the training tasks in both transfer learning settings. Then, we hypothesized that background planning would also suffer more in adapting to the subsequent test tasks in the adaptation setting. Finally, we hypothesized that, under certain assumptions, decision-time planning would perform better than background planning on the test tasks in the zero-shot setting.

Adaptation Scenario. In order to test the first two of these hypotheses, we compared the modern instantiations of the two planning styles on a sequential version of the RandDistShift (RDS) environment (Zhao et al., 2021). In this environment, the agent is first trained on training tasks with difficulty 0.35 (see Fig. 4e) and then it is left for adaptation to the test tasks with diffuculty 0.35 (see Fig. 4g). Results in Fig. 6f show that, similar to the planning & learning setting, the increasing usage of simulated experience can indeed slowdown, or even prevent (as in B(S)), background planning in reaching optimal (or close-to-optimal) performance on the training tasks, and that background planning indeed suffers more in adapting to the test tasks.

Zero-shot Scenario. And, in order to test the first and third hypotheses, we compared the two planning styles on the original RDS environment that was introduced by Zhao et al. (2021). In this environment, the agent is trained on training tasks with difficulty 0.35 (see Fig. 4e) and during the training process it is periodically evaluated on the test tasks with difficulties varying from 0.25 to 0.45 (see Fig. 4f-4h). Results are shown in Fig. 6g-6j. As can be seen in Fig. 6g, the increasing usage of simulated experience can again slowdown, or even prevent (as in B(S)), background planning in reaching optimal (or close-to-optimal) performance on the training tasks. We can also see in Fig. 6h-6j that decision-time planning indeed achieves significantly better zero-shot performance than background planning across all test tasks with varying difficulties.

6 Related Work

The abstract view of the two planning styles that we provide in this study is mostly related to the recent monograph of Bertsekas (2021) in which the recent successes of AlphaZero (Silver et al., 2018), a decision-time planning algorithm, are viewed through the lens of dynamic programming. However, we take a broader perspective and provide a unified view that encompasses both decision-time and background planning algorithms. Also, instead of assuming the availability of an exact model, we consider scenarios in which a model has to be learned by pure interaction with the environment. Another closely related study is the study of Hamrick et al. (2021) which informally relates MuZero (Schrittwieser et al., 2020), another decision-time planning algorithm, to various other decision-time and background planning algorithms in the literature. Our study can be viewed as a study that formalizes the relation between the two planning styles.

On the performance comparison side, there have also been benchmarking studies that empirically compare the performances of various decision-time and background planning algorithms on continuous control domains in the planning & learning setting (Wang et al., 2019), and on MiniGrid environments in specific transfer learning settings (Zhao et al., 2021). However, none of these studies provide a general understanding of when will one planning style perform better than the other. Also, rather than comparing the algorithms using the expected discounted return, these studies perform the comparison using the expected undiscounted return, and thus might be misleading in understanding the degree of optimality of the generated output policies.

Finally, our work also has connections to the studies of Jiang et al. (2015) and Arumugam et al. (2018) which provide upper bounds for the performance difference between policies generated as a result of planning with an exact and an estimated model. However, rather than providing upper bounds, in this study, we are interested in understanding which classes of models will allow for one planning style to perform better than

Table 1: Summary of how the simplest and modern instantiations two planning styles would compare against each other across different settings. Read from left to right in a top-down fashion.

Setting	Simplest Instantiations	Modern Instantiations
Planning & Learning	**1) Fairest Scenario** Background planning performs better when their model converges to a PRM (or a PXM) **Reason:** While decision-time planning corresponds to performing one-step PI, background planning corresponds to performing full PI **2) Common Scenario** Background planning performs better when its model converges to a PXM **Reason:** Due to the same reason as in the fairest scenario	**1) Simplified Scenario** The performance gap between the two planning styles that exists in the simplest instantiations reduces and it gradually closes after the models of both planning styles converge to PXMs **Reason:** Decision-time planning now corresponds to performing more than one-step PI and it makes use of an improving base policy **2) Original Scenario** Decision-time planning performs better even if the models of both planning styles converge to PXMs **Reason:** Background planning suffers from updating its value estimator with the harmful simulated experience generated by its parametric model, which can slowdown, or even prevent it in reaching optimal (or close-to-optimal) performance
Transfer Learning	**1) Adaptation Scenario** Similar to the common scenario of the planning & learning setting, background planning performs better when its model converges to a PXM **Reason:** Due to the same reason in the planning & learning setting	**1) Adaptation Scenario** Decision-time planning performs better even if the models of both planning styles converge to PXMs, both before and after the switch of tasks **Reason:** Before the switch of tasks, it is due to the same reason in the planning & learning setting After the switch of tasks, in addition to the reason in the planning & learning setting, it is also due to background planning's parametric model generating experience that resembles the training tasks until it adapts to the test tasks, which can lead to harmful updates to its value estimator **2) Zero-shot Scenario** Decision-time planning performs better if its parametric model is able to simulate at least a few time steps of the test tasks (i.e., if its parametric model generalizes to the test tasks) **Reason:** Decision-time planning would be able to improve upon the policy obtained during training by performing online planning

the other. Lastly, another related line of research is the recent studies of Grimm et al. (2020; 2021) which classify models according to how relevant they are for value-based planning. Although, we share the same overall idea that models should only be judged for how useful they are in the planning process, our work differs in that we classify models according to how useful they are in comparing the two planning styles.

7 Conclusion and Discussion

To summarize, we performed a unified analysis of decision-time and background planning and attempted to answer the following question:

> *Using the discounted return as the performance measure, under what conditions and in which settings will one planning style perform better than the other?*

In our analysis, we have tried to be as independent as possible from the specific implementation details of each planning style and tried to focus on the general working principles of them. Overall, our findings, summarized in Table 1, suggest that even though the simplest instantiations of decision-time planning do not perform as well as the simplest instantiations of background planning, the modern instantiations of it can perform on par or better than their background planning counterparts in both the planning & learning and transfer learning settings.

We note that the main purpose of this study was to contribute towards the goal of providing a *general understanding* of under what conditions and in which settings will one planning style perform better than the other through studying the generic algorithms in their corresponding classes, and *not* to provide a benchmark that compares state-of-the-art model-based RL algorithms across various settings and domains. We also note that even though providing practical insights is not the main goal of this study at the moment, we believe that our study can guide the community in improving the two planning styles in potentially interesting ways. For example, a possible improvement to the modern decision-time planning algorithm that is used in this study could be add a meta-level algorithm that controls how much the model should be unrolled at decision-time, and a possible improvement to the considered modern background planning algorithm could be to add a meta-level algorithm that balances the trade-off between the usage of real vs. simulated data in the updates of the value estimator. Finally, note that we were only interested in comparing the two planning styles in terms of the expected discounted return of their output policies. Though not the main focus of this study, other possible interesting comparison directions include comparing the two planning styles in terms of their sample efficiency and real-time performance.

References

Safa Alver and Doina Precup. Constructing a good behavior basis for transfer using generalized policy updates. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=7IWGzQ6gZ1D.

Ankesh Anand, Jacob C Walker, Yazhe Li, Eszter Vértes, Julian Schrittwieser, Sherjil Ozair, Theophane Weber, and Jessica B Hamrick. Procedural generalization by planning with self-supervised world models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=FmBegXJToY.

Dilip Arumugam, David Abel, Kavosh Asadi, Nakul Gopalan, Christopher Grimm, Jun Ki Lee, Lucas Lehnert, and Michael L Littman. Mitigating planner overfitting in model-based reinforcement learning. *arXiv preprint arXiv:1812.01129*, 2018.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

André Barreto, Diana Borsa, Shaobo Hou, Gheorghe Comanici, Eser Aygün, Philippe Hamel, Daniel Toyama, Jonathan Hunt, Shibl Mourad, David Silver, et al. The option keyboard combining skills in reinforcement learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 13052–13062, 2019.

André Barreto, Shaobo Hou, Diana Borsa, David Silver, and Doina Precup. Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences*, 117(48):30079–30087, 2020.

Dimitri P Bertsekas. *Rollout, policy iteration, and distributed reinforcement learning.* Athena Scientific, 2021.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming.* Athena Scientific, 1996.

Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2): 57–83, 2002.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid, 2018.

Christopher Grimm, Andre Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5541–5552. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/3bb585ea00014b0e3ebe4c6dd165a358-Paper.pdf.

Christopher Grimm, Andre Barreto, Gregory Farquhar, David Silver, and Satinder Singh. Proper value equivalence. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=aXbuWbta0V8.

Danijar Hafner, Timothy P Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0oabwyZbOu.

Jessica B Hamrick, Abram L. Friesen, Feryal Behbahani, Arthur Guez, Fabio Viola, Sims Witherspoon, Thomas Anthony, Lars Holger Buesing, Petar Veličković, and Theophane Weber. On the role of planning in model-based deep reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=IrM64DGB21.

Taher Jafferjee, Ehsan Imani, Erin Talvitie, Martha White, and Micheal Bowling. Hallucinating value: A pitfall of dyna-style planning with imperfect environment models. *arXiv preprint arXiv:2006.04363*, 2020.

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1181–1189. Citeseer, 2015.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pp. 282–293. Springer, 2006.

Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. 2002.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, pp. 216–224. Elsevier, 1990.

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Erik Talvitie. Model regularization for stable sample rollouts. In *UAI*, pp. 780–789, 2014.

Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

Gerald Tesauro. Td-gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation*, 6(2):215–219, 1994.

Gerald Tesauro and Gregory Galperin. On-line policy improvement using monte-carlo search. *Advances in Neural Information Processing Systems*, 9:1068–1074, 1996.

Hado P van Hasselt, Matteo Hessel, and John Aslanides. When to use parametric models in reinforcement learning? *Advances in Neural Information Processing Systems*, 32:14322–14333, 2019.

Harm Van Seijen, Hadi Nekoei, Evan Racah, and Sarath Chandar. The loca regret: A consistent metric to evaluate model-based behavior in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:6562–6572, 2020.

Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. Benchmarking model-based reinforcement learning. *arXiv preprint arXiv:1907.02057*, 2019.

Mingde Zhao, Zhen Liu, Sitao Luan, Shuyuan Zhang, Doina Precup, and Yoshua Bengio. A consciousness-inspired planning agent for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Łukasz Kaiser, Mohammad Babaeizadeh, Piotr Miłos, Błażej Osiński, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model based reinforcement learning for atari. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=S1xCPJHtDB`.

A Discussion on the Combined View of the Parametric and Non-Parametric Models

In order to be able to view the decision-time and background planning algorithms that perform planning with both a parametric and non-parametric model through our proposed unified framework, we view the two separate models of these algorithms as a single combined model. This becomes obvious for background planning algorithms if one notes that they perform planning with a batch of data that is jointly generated by both a parametric and non-parametric model (see e.g., line 20 in Alg. 6 in which ϕ_θ and Q_η are updated with a batch of data that is jointly generated by both $\bar{\bar{m}}_{p\omega}$ and $\bar{\bar{m}}_{np}$), which can be thought of performing planning with a batch of data that is generated by a single combined model. It also becomes obvious for decision-time planning algorithms if one notes that they perform planning by first performing search with a parametric model, and then by bootstrapping on the value estimates of a continually improving policy that is obtained by planning with a non-parametric model (see e.g., line 13 in Alg. 5 in which action selection is done with both $\bar{m}_{p\omega}$ and Q_η (which is obtained by planning with $\bar{m}_{np}$)), which can be thought of performing planning with a single combined model that is obtained by concatenating the parametric and non-parametric models.

B Discussion on the Choice of the Simplest Instantiations of the Two Planning Styles

As indicated in the main paper, for decision-time planning, we study the online Monte-Carlo planning (OMCP) algorithm of Tesauro & Galperin (1996), and, for background planning, we study the Dyna-Q algorithm of Sutton (1990; 1991). We choose these algorithms as they are the simplest instantiations in their corresponding classes and they are easy to analyze. In this study, as we are interested in scenarios where the model has to be learned from pure interaction, we consider a version of the OCMP algorithm in which a parametric model is learned from experience (see Alg. 1 for the pseudocode). Note that this is the only difference compared to the original version of the OMCP algorithm proposed in Tesauro & Galperin (1996). And, in order to make a fair comparison with this version of the OMCP algorithm, we consider a simplified version of the Dyna-Q algorithm (see Alg. 3 & 4 for the pseudocodes of the original and simplified versions, respectively). Compared to the original version of Dyna-Q, in this version, there are several minor differences:

- While planning, the agent can now sample states and actions that it has not observed or taken before. Note that this is also the case for the version of the OMCP algorithm considered in this study.
- Now, instead of using samples from both the environment and model, the agent updates its value estimator with samples only from the model. Note that the version of the OMCP algorithm considered in this study also makes use of only the model while performing planning.

Algorithm 1 Tabular Online Monte-Carlo Planning (Tesauro & Galperin, 1996) with an Adaptable Model

```
Initialize π^i ∈ Π as a random policy
Initialize m̄(s, a) ∀s ∈ S & ∀a ∈ A
n_r ← number of episodes to perform rollouts
while m̄ has not converged do
    S ← reset environment
    while not done do
        A ← ε-greedy(MC_rollout(S, m̄, n_r, π^i))
        R, S′, done ← environment(A)
        Update m̄(S, A) with R, S′, done
        S ← S′
    end while
end while
Return m̄(s, a)
```

Algorithm 2 Tabular Exhaustive Search (Campbell et al., 2002) with an Adaptable Model

Initialize $\bar{m}(s, a)$ $\forall s \in \mathcal{S}$ & $\forall a \in \mathcal{A}$
$h \leftarrow$ search heuristic
while $\bar{m}$ has not converged **do**
    $S \leftarrow$ reset environment
    while not done **do**
        $A \leftarrow \epsilon$-greedy(exhaustive_tree_search$(S, \bar{m}, h)$)
        $R, S', \text{done} \leftarrow \text{environment}(A)$
        Update $\bar{m}(S, A)$ with R, S', done
        $S \leftarrow S'$
    end while
end while
Return $\bar{m}(s, a)$

Algorithm 3 General Tabular Dyna-Q (Sutton, 1990; 1991)

Initialize $Q(s, a)$ $\forall s \in \mathcal{S}$ & $\forall a \in \mathcal{A}$
Initialize $\bar{\bar{m}}(s, a)$ $\forall s \in \mathcal{S}$ & $\forall a \in \mathcal{A}$
$\mathcal{SA}_{\text{prev}} \leftarrow \{\}$
$n_p \leftarrow$ number of time steps to perform planning
while Q and $\bar{\bar{m}}$ has not converged **do**
    $S \leftarrow$ reset environment
    while not done **do**
        $A \leftarrow \epsilon$-greedy$(Q(S, \cdot))$
        $R, S', \text{done} \leftarrow \text{environment}(A)$
        $\mathcal{SA}_{\text{prev}} \leftarrow \mathcal{SA}_{\text{prev}} + \{(S, A)\}$
        Update $Q(S, A)$ with R, S', done
        Update $\bar{\bar{m}}(S, A)$ with R, S', done
        $i \leftarrow 0$
        while $i < n_p$ **do**
            $S_{\bar{\bar{m}}}, A_{\bar{\bar{m}}} \leftarrow$ sample from $\mathcal{SA}_{\text{prev}}$
            $R_{\bar{\bar{m}}}, S'_{\bar{\bar{m}}}, \text{done}_{\bar{\bar{m}}} \leftarrow \bar{\bar{m}}(S_{\bar{\bar{m}}}, A_{\bar{\bar{m}}})$
            Update $Q(S_{\bar{\bar{m}}}, A_{\bar{\bar{m}}})$ with $R_{\bar{\bar{m}}}$, $S'_{\bar{\bar{m}}}$, $\text{done}_{\bar{\bar{m}}}$
            $i \leftarrow i + 1$
        end while
        $S \leftarrow S'$
    end while
end while
Return $Q(s, a)$

C Proofs

Proposition 1. *Let $m \in \mathcal{M}$ be a PRM of m^* with respect to $\Pi = \{\pi^r_m, \pi^{ce}_m\} \subseteq \mathbb{\Pi}$ and J. Then, $J^{\pi^{ce}_m}_{m^*} \geq J^{\pi^r_m}_{m^*}$.*

Proof. This result directly follows from Defn. 2 & 5. Recall that, according to Defn. 5, given a $\pi^b \in \mathbb{\Pi}$, π^r_m and π^{ce}_m are the policies that are obtained after performing one-step PI and full PI in model m, respectively. Thus, we have $J^{\pi^r_m}_m \leq J^{\pi^{ce}_m}_m$ (Bertsekas & Tsitsiklis, 1996), which, by Defn. 2, implies $J^{\pi^{ce}_m}_{m^*} \geq J^{\pi^r_m}_{m^*}$. □

Proposition 2. *Let $\bar{m} \in \mathcal{M}$ be any model of m^* with respect to $\bar{\Pi} = \{\pi^r_{\bar{m}}, \pi^{ce}_{\bar{m}}\} \subseteq \mathbb{\Pi}$ and J, and let $\bar{\bar{m}} \in \mathcal{M}$ be a PXM of m^* with respect to $\bar{\bar{\Pi}} = \{\pi^r_{\bar{\bar{m}}}, \pi^{ce}_{\bar{\bar{m}}}\} \subseteq \mathbb{\Pi}$ and J. Then, $J^{\pi^{ce}_{\bar{\bar{m}}}}_{m^*} \geq J^{\pi^r_{\bar{m}}}_{m^*}$.*

Proof. This result directly follows from Defn. 4 & 5. Recall that, according to Defn. 5, given a $\pi^b \in \mathbb{\Pi}$, $\pi^{ce}_{\bar{\bar{m}}}$ is the policy that is obtained after performing full PI in model $\bar{\bar{m}}$. Thus, $\pi^{ce}_{\bar{\bar{m}}}$ is one of the optimal

Algorithm 4 Tabular Dyna-Q of Interest

Initialize $Q(s, a)$ $\forall s \in \mathcal{S}$ & $\forall a \in \mathcal{A}$
Initialize $\bar{\bar{m}}(s, a)$ $\forall s \in \mathcal{S}$ & $\forall a \in \mathcal{A}$
$n_p \leftarrow$ number of time steps to perform planning
while Q and $\bar{\bar{m}}$ has not converged **do**
    $S \leftarrow$ reset environment
    while not done **do**
        $A \leftarrow \epsilon$-greedy$(Q(S, \cdot))$
        $R, S', \text{done} \leftarrow \text{environment}(A)$
        Update $\bar{\bar{m}}(S, A)$ with R, S', done
        $i \leftarrow 0$
        while $i < n_p$ **do**
            $S_{\bar{\bar{m}}}, A_{\bar{\bar{m}}} \leftarrow$ sample from $\mathcal{S} \times \mathcal{A}$
            $R_{\bar{\bar{m}}}, S'_{\bar{\bar{m}}}, \text{done}_{\bar{\bar{m}}} \leftarrow \bar{\bar{m}}(S_{\bar{\bar{m}}}, A_{\bar{\bar{m}}})$
            Update $Q(S_{\bar{\bar{m}}}, A_{\bar{\bar{m}}})$ with $R_{\bar{\bar{m}}}$, $S'_{\bar{\bar{m}}}$, $\text{done}_{\bar{\bar{m}}}$
            $i \leftarrow i + 1$
        end while
        $S \leftarrow S'$
    end while
end while
Return $Q(s, a)$

policies of model $\bar{\bar{m}}$ (Bertsekas & Tsitsiklis, 1996), which, by Defn. 4, implies $J^{\pi^{ce}_{\bar{\bar{m}}}}_{m^*} = \max_{\pi \in \Pi} J^{\pi}_{m^*}$ and thus $J^{\pi^{ce}_{\bar{\bar{m}}}}_{m^*} \geq J^{\pi}_{m^*} \forall \pi \in \Pi$. This in turn implies $J^{\pi^{ce}_{\bar{\bar{m}}}}_{m^*} \geq J^{\pi^{r}_{\bar{\bar{m}}}}_{m^*}$. □

D Discussion on the Choice of the Modern Instantiations of the Two Planning Styles

As indicated in the main paper, we study the decision-time and background planning algorithms in Zhao et al. (2021). More specifically, for decision-time planning, we study the "UP" algorithm, and, for background planning, we study the "Dyna" algorithm in Zhao et al. (2021). Note that these two algorithms do not employ the "bottleneck mechanism" introduced in Zhao et al. (2021). We choose these algorithms as they are reflective of many of the properties of their state-of-the-art counterparts (MuZero (Schrittwieser et al., 2020) for decision-time planning and SimPLe (Łukasz Kaiser et al., 2020) / DreamerV2 (Hafner et al., 2021) for background planning) and their code is publicly available[5]. The pseudocodes of these algorithms are presented in Alg. 5 & 6, respectively. Some of the important similarities and differences between these algorithms and their state-of-the-art counterparts are as follows:

1. Similarities and differences between the decision-time planning algorithm in Zhao et al. (2021) and MuZero (Schrittwieser et al., 2020):
    - Similar to MuZero, the decision-time planning algorithm in Zhao et al. (2021) also performs planning with both a parametric and non-parametric model.
    - Similar to MuZero, the decision-time planning algorithm in Zhao et al. (2021) also represents its parametric model using neural networks.
    - Similar to MuZero, the decision-time planning algorithm in Zhao et al. (2021) also learns a parametric model through pure interaction with the environment. However, rather than unrolling the model for several time steps and training it with the sum of the policy, value and reward losses as in MuZero, it unrolls the model for only a single time step and trains it with the sum of the value, dynamics, reward and termination losses. See the total loss term in Sec. 3 of Zhao et al. (2021).

[5]See `https://github.com/mila-iqia/Conscious-Planning` for the publicly available code.

- Lastly, similar to MuZero, the decision-time planning algorithm in Zhao et al. (2021) selects actions by directly bootstrapping on the value estimates of a continually improving policy (without performing any rollouts), which is obtained by planning with a non-parametric model, after performing some amount search with a parametric model. However, rather than performing the search using Monte-Carlo Tree Search (Kocsis & Szepesvári, 2006) as in MuZero, it uses best-first search during training and random search during evaluation. See App. H of Zhao et al. (2021) for the details of the search procedures.

2. Similarities and differences between the background planning algorithm in Zhao et al. (2021) and SimPLe (Łukasz Kaiser et al., 2020) / DreamerV2 (Hafner et al., 2021):
   - Similar to SimPLe / DreamerV2, the background planning algorithm in Zhao et al. (2021) also performs planning with a parametric model. Additionally, it also performs planning with a non-parametric model (a replay buffer).
   - Similar to SimPLe / DreamerV2, the background planning algorithm in Zhao et al. (2021) also represents its parametric model using neural networks and it also updates its value estimator using the simulated data generated with this model.
   - Similar to SimPLe / DreamerV2, the background planning algorithm in Zhao et al. (2021) also learns a parametric model through pure interaction with the environment. However, rather than performing planning with this model after allowing for an initial kickstarting period as in SimPLe / DreamerV2, which is referred to as a "world model" learning period, for a fair comparison with the decision-time planning algorithm, it starts to perform planning right at the beginning of the model learning process.
   - Lastly, similar to SimPLe / DreamerV2, the background planning algorithm in Zhao et al. (2021) selects actions by simply querying its value estimator.

Note that even though there are some differences between the planning algorithms in Zhao et al. (2021) and their state-of-the-art counterparts, except for the kickstarting period in state-of-the-art background planning algorithms, these are just minor implementation details that would have no impact on the conclusions of this study. Although the kickstarting period can mitigate the harmful simulated data problem to some degree, or even prevent it if the period is sufficiently long, allowing for it would definitely prevent a fair comparison with the decision-time planning algorithm in Zhao et al. (2021). This is why we did not allow for this period in our experiments.

E Experimental Details

In this section, we provide the details of the experiments that are performed in Sec. 5. This also includes the implementation details of the simplest and modern instantiations of the two planning styles considered in this study.

E.1 Details of the Simplest Instantiation Experiments

In all of the simplest instantiation experiments, we have calculated the performance with a discount factor of 0.9.

Environment & Models. All of the experiments in Sec. 5.1 are performed on the Simple Gridworld environment. Here, the agent spawns in state S and has to navigate to the goal state depicted by G. At each time step, the agent receives an (x, y) pair, indicating its position, and based on this, selects an action that moves it to one of the four neighboring cells with a slip probability of 0.05. The agent receives a negative reward that is linearly proportional to its distance from G and a reward of +10 if it reaches G (see Fig. E.2a). The agent-environment interaction lasts for a maximum of 100 time steps, after which the episode terminates with a reward of 0 if the agent was not able to reach the goal state G.

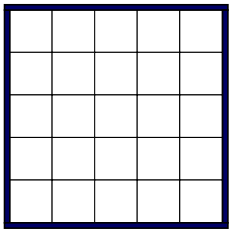


Figure E.1: The form of state aggregation used in this study.

Algorithm 5 The Decision-Time Planning Algorithm in Zhao et al. (2021)

Initialize the parameters θ, η & ω of $\phi_\theta : \mathcal{O}_E \rightarrow \mathcal{S}_A$, $Q_\eta : \mathcal{S}_A \times \mathcal{A}_E \rightarrow \mathbb{R}$ & $\bar{m}_{p\omega} = (p_\omega, r_\omega, d_\omega)$
Initialize the replay buffer $\bar{m}_{np} \leftarrow \{\}$
$N_{ple} \leftarrow$ number of episodes to perform planning and learning
$N_{rbt} \leftarrow$ number of samples that the replay buffer must hold to perform planning and learning
$n_s \leftarrow$ number of time steps to perform search
$n_{bs} \leftarrow$ number of samples to sample from $\bar{m}_{np}$
$h \leftarrow$ search heuristic
$S \leftarrow$ replay buffer sampling strategy
$i \leftarrow 0$
while $i < N_{ple}$ **do**
    $O \leftarrow$ reset environment
    while not done **do**
        $A \leftarrow \epsilon$-greedy(tree_search_with_bootstrapping$(\phi_\theta(O), \bar{m}_{p\omega}, Q_\eta, n_s, h)$)
        $R, O', \text{done} \leftarrow \text{environment}(A)$
        $\bar{m}_{np} \leftarrow \bar{m}_{np} + \{(O, A, R, O', \text{done})\}$
        if $|\bar{m}_{np}| \geq N_{rbt}$ **then**
            $\mathcal{B}_{np} \leftarrow$ sample_batch$(\bar{m}_{np}, n_{bs}, S)$
            Update ϕ_θ, Q_η & $\bar{m}_{p\omega}$ with $\mathcal{B}_{np}$
        end if
        $O \leftarrow O'$
    end while
    $i \leftarrow i + 1$
end while
Return ϕ_θ, Q_η & $\bar{m}_{p\omega}$

-4.5	-4	-3.5	-3	-2.5	-2	-1.5	-1	-0.5	+10
-5	-4.5	-4	-3.5	-3	-2.5	-2	-1.5	-1	-0.5
-5.5	-5	-4.5	-4	-3.5	-3	-2.5	-2	-1.5	-1
-6	-5.5	-5	-4.5	-4	-3.5	-3	-2.5	-2	-1.5
-6.5	-6	-5.5	-5	-4.5	-4	-3.5	-3	-2.5	-2
-7	-6.5	-6	-5.5	-5	-4.5	-4	-3.5	-3	-2.5
-7.5	-7	-6.5	-6	-5.5	-5	-4.5	-4	-3.5	-3
-8	-7.5	-7	-6.5	-6	-5.5	-5	-4.5	-4	-3.5
-8.5	-8	-7.5	-7	-6.5	-6	-5.5	-5	-4.5	-4
-9	-8.5	-8	-7.5	-7	-6.5	-6	-5.5	-5	-4.5

(a) Simple Gridworld Rewards

0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	+10	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

(b) m_8 Rewards

-4.5	-5	-5.5	-6	-6.5	-7	-7.5	-8	-8.5	-9
-4	-4.5	-5	-5.5	-6	-6.5	-7	-7.5	-8	-8.5
-3.5	-4	-4.5	-5	-5.5	-6	-6.5	-7	-7.5	-8
-3	-3.5	-4	-4.5	-5	-5.5	-6	-6.5	-7	-7.5
-2.5	-3	-3.5	-4	-4.5	-5	-5.5	-6	-6.5	-7
-2	-2.5	-3	-3.5	-4	-4.5	-5	-5.5	-6	-6.5
-1.5	-2	-2.5	-3	-3.5	-4	-4.5	-5	-5.5	-6
-1	-1.5	-2	-2.5	-3	-3.5	-4	-4.5	-5	-5.5
-0.5	-1	-1.5	-2	-2.5	-3	-3.5	-4	-4.5	-5
+10	-0.5	-1	-1.5	-2	-2.5	-3	-3.5	-4	-4.5

(c) Test Task Rewards

Figure E.2: Reward functions of (a) the Simple Gridworld environment, (b) the m_8 model and (c) the test task.

- **Planning & Learning Setting.** (Fairest Scenario) For the experiments in the fairest scenario of the planning & learning setting, m was trained to sequentially converge to a set of PRMs in which the agent receives a reward of $+10$ if it reaches the goal state and a reward of 0 elsewhere. For example, see the reward function of model m_8 in Fig. E.2b. Note that these models have the same transition distribution and initial state distribution with the Simple Gridworld environment. (Common Scenario) For the experiments in the practical scenarios of the planning & learning setting, we have assumed that the agent already has access to the transition distribution and initial state distribution of the environment, and only has to learn the reward function.

- **Transfer Learning Setting.** Finally, for the experiments in the transfer learning setting, we considered a test task with a reward function as in Fig. E.2c, which is a transposed version of the training task's reward function (see Fig. E.2a). Note again that we have assumed that the agent already has access to the transition distribution and initial state distribution of the environment, and only has to learn the reward function.

Algorithm 6 The Background Planning Algorithm in Zhao et al. (2021)

Initialize the parameters θ, η & ω of $\phi_\theta : \mathcal{O}_E \rightarrow \mathcal{S}_A$, $Q_\eta : \mathcal{S}_A \times \mathcal{A}_E \rightarrow \mathbb{R}$ & $\bar{\bar{m}}_{p\omega} = (p_\omega, r_\omega, d_\omega)$
Initialize the replay buffer $\bar{\bar{m}}_{np} \leftarrow \{\}$ and the imagined replay buffer $\bar{\bar{m}}_{inp} \leftarrow \{\}$
$N_{ple} \leftarrow$ number of episodes to perform planning and learning
$N_{rbt} \leftarrow$ number of samples that the replay buffer must hold to perform planning and learning
$n_{ibs} \leftarrow$ number of samples to sample from $\bar{\bar{m}}_{inp}$
$n_{bs} \leftarrow$ number of samples to sample from $\bar{\bar{m}}_{np}$
$S \leftarrow$ replay buffer sampling strategy
$i \leftarrow 0$
while $i < N_{ple}$ **do**
    $O \leftarrow$ reset environment
    while not done **do**
        $A \leftarrow \epsilon$-greedy$(Q_\eta(\phi_\theta(O), \cdot))$
        $R, O', \text{done} \leftarrow \text{environment}(A)$
        $\bar{\bar{m}}_{np} \leftarrow \bar{\bar{m}}_{np} + \{(O, A, R, O', \text{done})\}$
        $\bar{\bar{m}}_{inp} \leftarrow \bar{\bar{m}}_{inp} + \{(\phi_\theta(O), A)\}$
        if $|\bar{\bar{m}}_{np}| \geq N_{rbt}$ **then**
            $\mathcal{B}_{inp} \leftarrow \text{sample_batch}(\bar{\bar{m}}_{inp}, n_{ibs}, S)$
            $\mathcal{B}_p \leftarrow \mathcal{B}_{inp} + \bar{\bar{m}}_{p\omega}(\mathcal{B}_{inp})$
            $\mathcal{B}_{np} \leftarrow \text{sample_batch}(\bar{\bar{m}}_{np}, n_{bs}, S)$
            Update ϕ_θ & Q_η with $\mathcal{B}_{np} + \mathcal{B}_p$
            Update ϕ_θ & $\bar{\bar{m}}_{p\omega}$ with $\mathcal{B}_{np}$
        end if
        $O \leftarrow O'$
    end while
    $i \leftarrow i + 1$
end while
Return ϕ_θ & Q_η

Implementation Details of the Simplest Instantiations. For our simplest instantiation experiments, we considered specific versions of the OMCP algorithm of Tesauro & Galperin (1996) and the Dyna-Q algorithm of Sutton (1990; 1991). The pseudocodes of these algorithms are presented in Alg. 1 & 4, respectively, and the details of them are provided in Table E.1 & E.2, respectively.

Table E.1: Details and hyperparameters of Alg. 1.

π^i	a deterministic random policy
$\bar{m}$	a tabular model (initialized as a hand-designed PNM, see Fig. **??**)
n_r	50
ϵ	linearly decays from 1.0 to 0.0 over the first 20 episodes
State Aggregation	an aggregation of the form in Fig. E.1

Table E.2: Details and hyperparameters of Alg. 4.

Q	a tabular value function (initialized as zero $\forall s \in \mathcal{S}$ and $\forall a \in \mathcal{A}$)
$\bar{\bar{m}}$	a tabular model (initialized as a hand-designed PNM, see Fig. **??**)
n_p	100
ϵ	linearly decays from 1.0 to 0.0 over the first 20 episodes
State Aggregation	an aggregation of the form in Fig. E.1

Allowed Computational Budget at Decision Time. For all of our experiments we limit the n_r parameter in Alg. 1 (the number of episodes to perform rollouts at decision time) to 50 and we limit the n_p parameter in Alg. 4 (the number of time steps to to update the value estimator at decision time) to 500.

E.2 Details of the Modern Instantiation Experiments

In all of the modern instantiation experiments on the Simple Gridworld environment, we have again calculated the performance with a discount factor of 0.9, and in all of the modern instantiation experiments on the MiniGrid environments, we have calculated the performance with a discount factor of 0.99.

Environments & Models. In Sec. 5.2, part of our experiments were performed on the Simple Gridworld environment and part of them were performed on MiniGrid environments. The details of these environments and their corresponding models are as follows:

- **Simple Gridworld Environment.** To learn about the Simple Gridworld environment, we refer the reader to Sec. E.1 as we have used the same environment in the modern instantiation experiments as well.
  - **Planning & Learning Setting.** To learn about the models of both planning styles, we also refer the reader to the planning & learning setting of Sec. E.1 as have used the same models in the modern instantiation experiments as well.
- **MiniGrid Environments.** In the MiniGrid environments, the agent (depicted in red) has to navigate to the green goal cell while avoiding the orange lava cells (if there are any). At each time step, the agent receives a grid-based observation that contains its own position and the positions of the goal, wall and lava cells, and based on this, selects an action that either turns it left or right, or steps it forward. If the agent steps on a lava cell, the episode terminates with no reward, and if it reaches the goal cell, the episode terminates with a reward of $+1$.[6] More on the details of these environments can be found in Chevalier-Boisvert et al. (2018).
  - **Planning & Learning Setting.** For the planning & learning setting, we performed experiments on the Empty 10x10, FourRooms, SimpleCrossingS9N1 and LavaCrossingS9N1 environments (see Fig. 4a-4d). While the last two of these environments already pre-exist in MiniGrid, the first two of them are manually built environments. Specifically, the Empty 10x10 environment is obtained by expanding the Empty 8x8 environment and the 10x10 FourRooms environment is obtained by contracting the 16x16 FourRooms environment in MiniGrid.
  - **Transfer Learning Setting.** For the transfer learning setting, we performed experiments on the sequential and regular versions of the RandDistShift (RDS) environment considered in Zhao et al. (2021) (see Fig. 4e-4h). We have already presented the details of these environments in the main article. To learn more about the details, we refer the reader to Zhao et al. (2021).

  Finally, note that, as opposed to the experiments on Simple Gridworld environment, for both the planning & learning and transfer learning settings, we did not enforce any kind of structure on the models of the agent, and just initialized them randomly. We also note that, in our experiments with RDS Sequential, we reinitialized the non-parametric models (replay buffers) of both planning styles after the tasks switch from the training tasks to the test tasks.

Implementation Details of the Modern Instantiations. For our modern instantiation experiments, we considered the decision-time and background planning algorithms in Zhao et al. (2021) (see Sec. D), whose pseudocodes are presented in Alg. 5 & 6, respectively. The details of these algorithms are provided in Table E.3 & E.4, respectively. For more details such as the neural network architectures, replay buffer sizes, learning rates, exact details of the tree search ..., we refer the reader to the publicly available code and the supplementary material of Zhao et al. (2021).

Note that the publicly available code of Zhao et al. (2021) only contains background planning algorithms in which the model is a model over the observations (and not the states), which for some reason causes the background planning algorithm of interest to perform very poorly (see the plots in Zhao et al. (2021)). For this reason and in order to make a fair comparison with decision-time planning, we have implemented a version of

[6] Note that this is not the original reward function of MiniGrid environments. In the original version, the agent receives a reward of $+1 - 0.9(t/T)$, where t is the number of time steps taken to reach the goal cell and T is the maximum episode length, if it reaches the goal. We modified the reward function in order to obtain more intuitive results.

Table E.3: Details and hyperparameters of Alg. 5.

ϕ_θ	MiniGrid bag of words feature extractor
Q_η	regular neural network (planing & learning setting), neural network with attention for set-based representations (transfer learning setting)
$\bar{m}_{p\omega}$	regular neural network (planning & learning setting), neural network with attention (for set-based representations) (transfer learning setting) (bottleneck mechanism is disabled for both of the settings)
N_{ple}	50M
N_{rbt}	50k
n_s	1 (DT(1)), 5 (DT(5)), 15 (DT(15))
n_{bs}	128 (planning & learning setting), 64 (transfer learning setting)
h	best-first search (training), random search (evaluation)
S	random sampling
ϵ	linearly decays from 1.0 to 0.0 over the first 1M time steps

Table E.4: Details and hyperparameters of Alg. 6.

ϕ_θ	MiniGrid bag of words feature extractor
Q_η	regular neural network (planning & learning setting), neural network with attention (for set-based representations) (transfer learning setting)
$\bar{\bar{m}}_{p\omega}$	regular neural network (planning & learning setting), neural network with attention (for set-based representations) (transfer learning setting) (bottleneck mechanism is disabled for both of the settings)
N_{ple}	50M
N_{rbt}	50k
(n_{ibs}, n_{bs})	$(0, 128)$ (B(R)), $(128, 128)$ (B(R+S)), $(128, 0)$ (B(S)) (planning & learning setting), $(0, 64)$ (B(R)), $(64, 64)$ (B(R+S)), $(64, 0)$ (B(S)) (transfer learning setting)
S	random sampling
ϵ	linearly decays from 1.0 to 0.0 over the first 1M time steps

the background planning algorithm in which the model is a model over the states and we performed all of our experiments with this version of the algorithm. Also note that, while we have used regular representations in our planning & learning experiments, in order to deal with the large number of tasks, we have made use of set-based representations in our transfer learning experiments. See Zhao et al. (2021) for the details of this representation.

Additionally, we also performed experiments with simplified tabular versions of the modern instantiations of the two planning styles, whose pseudocodes are presented in Alg. 7 & 8, respectively. The details of these algorithms are provided in Table E.5 & E.6, respectively.

Table E.5: Details and hyperparameters of Alg. 7.

Q	a tabular value function (initialized as zero $\forall s \in \mathcal{S}$ and $\forall a \in \mathcal{A}$)
$\bar{m}_p$	a tabular model (initialized as a hand-designed PNM (see Fig. **??**))
n_s	$\lvert\mathcal{A}\rvert$
h	breadth-first search
ϵ	linearly decays from 1.0 to 0.0 over the first 20 episodes

Table E.6: Details and hyperparameters of Alg. 8.

Q	a tabular value function (initialized as zero $\forall s \in \mathcal{S}$ and $\forall a \in \mathcal{A}$)
$\bar{\bar{m}}_p$	a tabular parametric model (initialized as a hand-designed PNM (see Fig. **??**))
n_p	50
ϵ	linearly decays from 1.0 to 0.0 over the first 20 episodes

Algorithm 7 Modernized Version of Tabular OMCP with both a Parametric and Non-Parametric Model

Initialize $Q(s, a)$ $\forall s \in \mathcal{S}$ & $\forall a \in \mathcal{A}$
Initialize $\bar{m}_p(s, a)$ $\forall s \in \mathcal{S}$ & $\forall a \in \mathcal{A}$
Initialize the replay buffer $\bar{m}_{np} \leftarrow \{\}$
$n_s \leftarrow$ number of time steps to perform search
$h \leftarrow$ search heuristic
while $\bar{m}_p$ and $\bar{m}_{np}$ has not converged **do**
    $S \leftarrow$ reset environment
    while not done **do**
        $A \leftarrow \epsilon$-greedy(tree_search_with_bootstrapping$(S, \bar{m}_p, Q, n_s, h)$)
        $R, S', \text{done} \leftarrow \text{environment}(A)$
        $\bar{m}_{np} \leftarrow \bar{m}_{np} + \{(S, A, R, S', \text{done})\}$
        $S_{\bar{m}_{np}}, A_{\bar{m}_{np}}, R_{\bar{m}_{np}}, S'_{\bar{m}_{np}}, \text{done}_{\bar{m}_{np}} \leftarrow$ sample from $\bar{m}_{np}$
        Update Q & $\bar{m}_p$ with $S_{\bar{m}_{np}}, A_{\bar{m}_{np}}, R_{\bar{m}_{np}}, S'_{\bar{m}_{np}}, \text{done}_{\bar{m}_{np}}$
        $S \leftarrow S'$
    end while
end while
Return Q & $\bar{m}_p(s, a)$

Algorithm 8 Modernized Version of Tabular Dyna-Q of Interest with both a Parametric and Non-Parametric Model

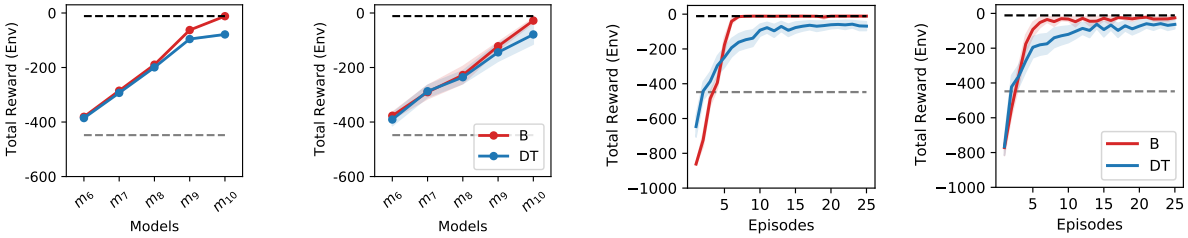
Initialize $Q(s, a)$ $\forall s \in \mathcal{S}$ & $\forall a \in \mathcal{A}$
Initialize $\bar{\bar{m}}_p(s, a)$ $\forall s \in \mathcal{S}$ & $\forall a \in \mathcal{A}$
Initialize the replay buffer $\bar{\bar{m}}_{np} \leftarrow \{\}$
$n_p \leftarrow$ number of time steps to perform planning
while Q, $\bar{\bar{m}}_p$ and $\bar{\bar{m}}_{np}$ has not converged **do**
    $S \leftarrow$ reset environment
    while not done **do**
        $A \leftarrow \epsilon$-greedy$(Q(S, \cdot))$
        $R, S', \text{done} \leftarrow \text{environment}(A)$
        Update $\bar{\bar{m}}_p(S, A)$ with R, S', done
        $\bar{\bar{m}}_{np} \leftarrow \bar{\bar{m}}_{np} + \{(S, A, R, S', \text{done})\}$
        $i \leftarrow 0$
        while $i < n_p$ **do**
            $S_{\bar{\bar{m}}_p}, A_{\bar{\bar{m}}_p} \leftarrow$ sample from $\mathcal{S} \times \mathcal{A}$
            $R_{\bar{\bar{m}}_p}, S'_{\bar{\bar{m}}_p}, \text{done}_{\bar{\bar{m}}_p} \leftarrow \bar{\bar{m}}_p(S_{\bar{\bar{m}}_p}, A_{\bar{\bar{m}}_p})$
            Update $Q(S_{\bar{\bar{m}}_p}, A_{\bar{\bar{m}}_p})$ with $R_{\bar{\bar{m}}_p}$, $S'_{\bar{\bar{m}}_p}$, $\text{done}_{\bar{\bar{m}}_p}$
            $S_{\bar{\bar{m}}_{np}}, A_{\bar{\bar{m}}_{np}}, R_{\bar{\bar{m}}_{np}}, S'_{\bar{\bar{m}}_{np}}, \text{done}_{\bar{\bar{m}}_{np}} \leftarrow$ sample from $\bar{\bar{m}}_{np}$
            Update $Q(S_{\bar{\bar{m}}_{np}}, A_{\bar{\bar{m}}_{np}})$ with $R_{\bar{\bar{m}}_{np}}$, $S'_{\bar{\bar{m}}_{np}}$, $\text{done}_{\bar{\bar{m}}_{np}}$
            $i \leftarrow i + 1$
        end while
        $S \leftarrow S'$
    end while
end while
Return $Q(s, a)$

Allowed Computational Budget at Decision Time. For all of our experiments we limit the n_s parameter in Alg. 5 (the number of steps to perform search at decision time) to 15 and we limit the n_{ibs} parameter (the number of samples to sample from the imagined replay buffer at decision time) in Alg. 6 to 128.

F Additional Results

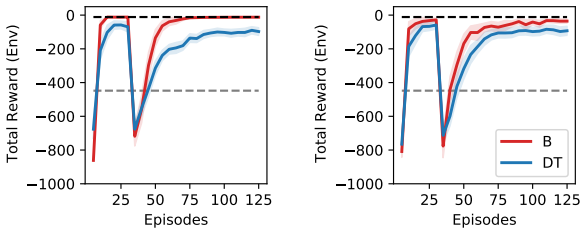
In this section, we provide total reward plots that are obtained by simply adding the rewards obtained by the agent throughout the episodes.

F.1 Experiments with Simplest Instantiations



(a) Tabular (b) State Aggregation (c) Tabular (d) State Aggregation

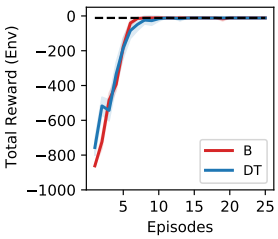
Figure F.1: The total reward obtained by the simplest instantiations of decision-time (DT) and background (B) planning on the Simple Gridworld environment (a, b) in the fairest scenario and (c, d) in the common scenario of the planning & learning setting with tabular and state aggregation value estimator representations. Black & gray dashed lines indicate total reward obtained by the optimal & random policies, respectively. Shaded regions are one standard error over 250 runs.



(a) Tabular (b) State Aggregation

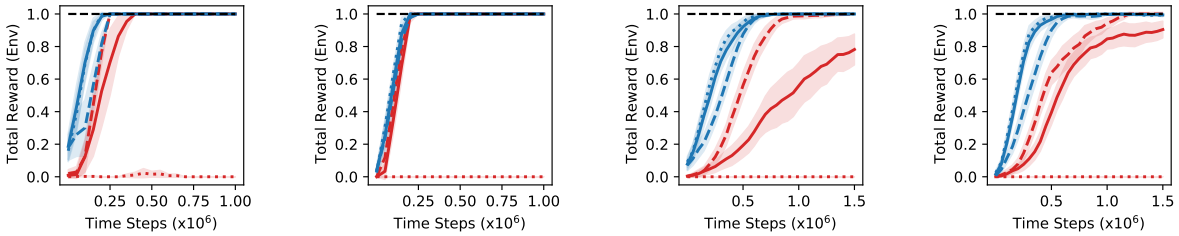
Figure F.2: The total reward obtained by the simplest instantiations of decision-time (DT) and background (B) planning on the Simple Gridworld environment in the transfer learning setting with tabular and state aggregation value estimator representations. Black & gray dashed lines indicate total reward obtained by the optimal & random policies, respectively. Shaded regions are one standard error over 250 runs.

F.2 Experiments with Modern Instantiations



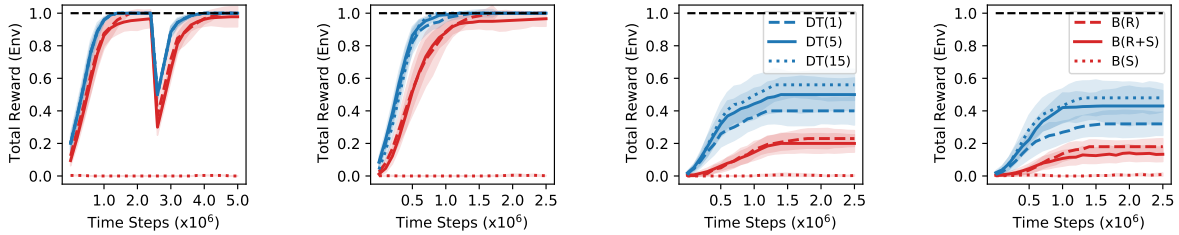
(a) Tabular

Figure F.3: The total reward obtained by the modern instantiations of decision-time (DT) and background (B) planning on the Simple Gridworld environment in the planning & learning setting with tabular value estimator representations. The black dashed line indicates the total reward obtained by the optimal policy. Shaded regions are one standard error over 250 runs.



(a) Empty 10x10 (b) FourRooms (c) SimpleCrossingS9N1 (d) LavaCrossingS9N1

Figure F.4: The total reward obtained by the modern instantiations of decision-time (DT) and background (B) planning in the planning & learning setting with neural network value estimator representations. The black dashed lines indicate the total reward obtained by the optimal policy in the corresponding environment. Shaded regions are one standard error over 100 runs.



(a) RDS Sequential (b) RDS Train (Diff: 0.35) (c) RDS Test (Diff: 0.35) (d) RDS Test (Diff: 0.45)

Figure F.5: The total reward obtained by the modern instantiations of decision-time (DT) and background (B) planning in the transfer learning setting with neural network value estimator representations. The black dashed lines indicate the total reward obtained by the optimal policy in the corresponding environment. Shaded regions are one standard error over 100 runs.