Online Planning with Lookahead Policies

Yonathan Efroni [*] [†] [‡] **Mohammad Ghavamzadeh** [§] **Shie Mannor** [‡] [¶]

Abstract

Real Time Dynamic Programming (RTDP) is an online algorithm based on Dynamic Programming (DP) that acts by 1-step greedy planning. Unlike DP, RTDP does not require access to the entire state space, i.e., it explicitly handles the exploration. This fact makes RTDP particularly appealing when the state space is large and it is not possible to update all states simultaneously. In this we devise a multi-step greedy RTDP algorithm, which we call h-RTDP, that replaces the 1-step greedy policy with a h-step lookahead policy. We analyze h-RTDP in its exact form and establish that increasing the lookahead horizon, h, results in an improved sample complexity, with the cost of additional computations. This is the first work that proves improved sample complexity as a result of *increasing* the lookahead horizon in online planning. We then analyze the performance of h-RTDP in three approximate settings: approximate model, approximate value updates, and approximate state representation. For these cases, we prove that the asymptotic performance of h-RTDP remains the same as that of a corresponding approximate DP algorithm, the best one can hope for without further assumptions on the approximation errors.

1 Introduction

Dynamic Programming (DP) algorithms return an optimal policy, given a model of the environment. Their convergence in the presence of lookahead policies [4, 13] and their performance in different approximate settings [4, 25, 27, 17, 1, 14] have been well-studied. Standard DP algorithms require simultaneous access to the *entire* state space at run time, and as such, cannot be used in practice when the number of states is too large. Real Time Dynamic Programming (RTDP) [3, 29] is a DP-based algorithm that mitigates the need to access all states simultaneously. Similarly to DP, RTDP updates are based on the Bellman operator, calculated by accessing the model of the environment. However, unlike DP, RTDP learns how to act by interacting with the environment. In each episode, RTDP interacts with the environment, acts according to the greedy action w.r.t. the Bellman operator, and samples a trajectory. RTDP is, therefore, an online planning algorithm.

Despite the popularity and simplicity of RTDP and its extensions [5, 6, 24, 8, 29, 22], precise characterization of its convergence was only recently established for finite-horizon MDPs [15]. While lookahead policies in RTDP are expected to improve the convergence in some of these scenarios, as they do for DP [4, 13], to the best of our knowledge, these questions have not been addressed in previous literature. Moreover, previous research haven't addressed the questions of how lookahead policies should be used in RTDP, nor studied RTDP's sensitivity to possible approximation errors. Such errors can arise due to a misspecified model, or exist in value function updates, when e.g., function approximation is used.

[*]Part of this work was done during an internship in Facebook AI Research
[†]Microsoft Research, New York, NY
[‡]Technion, Israel
[§]Google Research
[¶]Nvidia Research

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

In this paper, we initiate a comprehensive study of lookahead-policy based RTDP with approximation errors in *online planning*. We start by addressing the computational complexity of calculating lookahead policies and study its advantages in approximate settings. Lookahead policies can be computed naively by exhaustive search in $O(A^h)$ for deterministic environments or $O(A^{Sh})$ for stochastic environments. Since such an approach is infeasible, we offer in Section 3 an alternative approach for obtaining a lookahead policy with a computational cost that depends linearly on a natural measure: the total number of states reachable from a state in h time steps. The suggested approach is applicable both in deterministic and stochastic environments.

In Section 5, we introduce and analyze h-RTDP, a RTDP-based algorithm that replaces the 1-step greedy used in RTDP by a h-step lookahead policy. The analysis of h-RTDP reveals that the sample complexity is improved by increasing the lookahead horizon h. To the best of our knowledge, this is the first theoretical result that relates sample complexity to the lookahead horizon in online planning setting. In Section 6, we analyze h-RTDP in the presence of three types of approximation: when (i) an inexact model is used, instead of the true one, (ii) the value updates contain error, and finally (iii) approximate state abstraction is used. Interestingly, for approximate state abstraction, h-RTDP convergence and computational complexity depends on the size of the *abstract state space*.

In a broader context, this work shows that RTDP-like algorithms could be a good alternative to Monte Carlo tree search (MCTS) [7] algorithms, such as upper confidence trees (UCT) [21], an issue that was empirically investigated in [22]. We establish strong convergence guarantees for extensions of h-RTDP: under no assumption other than initial optimistic value, RTDP-like algorithms combined with lookahead policies converge in polynomial time to an optimal policy (see Table 1), and their approximations inherit the asymptotic performance of approximate DP (ADP). Unlike RTDP, MCTS acts by using a $\sqrt{\log N/N}$ bonus term instead of optimistic initialization. However, in general, its convergence can be quite poor, even worse than uniformly random sampling [9, 26].

2 Preliminaries

Finite Horizon MDPs. A finite-horizon MDP [4] with time-independent dynamics[6] is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, p, H)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces with cardinalities S and A, respectively, $r(s,a) \in [0,1]$ is the immediate reward of taking action a at state s, and $p(s'|s,a)$ is the probability of transitioning to state s' upon taking action a at state s. The initial state in each episode is arbitrarily chosen and $H \in \mathbb{N}$ is the MDP's horizon. For any $N \in \mathbb{N}$, denote $[N] := \{1, \dots, N\}$.

A deterministic policy $\pi : \mathcal{S} \times [H] \to \mathcal{A}$ is a mapping from states and time step indices to actions. We denote by $a_t := \pi_t(s)$ the action taken at time t at state s according to a policy π. The quality of a policy π from a state s at time t is measured by its value function, i.e., $V_t^\pi(s) := \mathbb{E}\big[\sum_{t'=t}^{H} r(s_{t'}, \pi_{t'}(s_{t'})) \mid s_t = s\big]$, where the expectation is over all the randomness in the environment. An optimal policy maximizes this value for all states $s \in \mathcal{S}$ and time steps $t \in [H]$, i.e., $V_t^*(s) := \max_\pi V_t^\pi(s)$, and satisfies the optimal Bellman equation,

$$\begin{aligned} V_t^*(s) = TV_{t+1}^*(s) &:= \max_a \big(r(s,a) + p(\cdot|s,a)V_{t+1}^*\big) \\ &= \max_a \mathbb{E}\big[r(s_1,a) + V_{t+1}^*(s_2) \mid s_1 = s\big]. \end{aligned} \tag{1}$$

By repeatedly applying the optimal Bellman operator T, for any $h \in [H]$, we have

$$\begin{aligned} V_t^*(s) = T^h V_{t+h}^*(s) &= \max_a \big(r(s,a) + p(\cdot|s,a)T^{h-1}V_{t+h}^*\big) \\ &= \max_{\pi_t,\dots,\pi_{t+h-1}} \mathbb{E}\Big[\sum_{t'=1}^{h} r(s_{t'}, \pi_{t+t'-1}(s_{t'})) + V_{t+h}^*(s_{h+1}) \mid s_1 = s\Big]. \end{aligned} \tag{2}$$

We refer to T^h as the h-step optimal Bellman operator. Similar Bellman recursion is defined for the value of a given policy, π, i.e., V^π, as $V_t^\pi(s) = T_\pi^h V_{t+h}^\pi(s) := r(s,\pi_t(s)) + p(\cdot|s,\pi_t(s))T_\pi^{h-1}V_{t+h}^\pi$, where T_π^h is the h-step Bellman operator of policy π.

h-Lookahead Policy. An h-lookahead policy w.r.t. a value function $V \in \mathbb{R}^S$ returns the optimal first action in an h-horizon MDP. For a state $s \in \mathcal{S}$, it returns

[6] The results can also be applied to time-dependent MDPs, however, the notations will be more involved.

$$a_h(s) \in \arg\max_a \big(r(s,a) + p(\cdot|s,a)T^{h-1}V\big)$$
$$= \arg\max_{\pi_1(s)} \max_{\pi_2,\ldots,\pi_h} \mathbb{E}\Big[\sum_{t=1}^{h} r(s_t,\pi_t(s_t)) + V(s_{h+1})|s_1 = s\Big]. \quad (3)$$

We can see V represent our 'prior-knowledge' of the problem. For example, it is possible to show [4] that if V is close to V^*, then the value of a h-lookahead policy w.r.t. V is close to V^*.

For a state $s \in \mathcal{S}$ and a number of time steps $h \in [H]$, we define the set of reachable states from s in h steps as $\mathcal{S}_h(s) = \{s' \mid \exists\pi : p^\pi(s_{h+1} = s' \mid s_1 = s, \pi) > 0\}$, and denote by $S_h(s)$ its cardinality. We define the set of reachable states from s in up to h steps as $\mathcal{S}_h^{Tot}(s) := \cup_{t=1}^h \mathcal{S}_t(s)$, its cardinality as $S_h^{Tot}(s) := \sum_{t=1}^h S_t(s)$, and the maximum of this quantity over the entire state space as $S_h^{Tot} = \max_s S_h^{Tot}(s)$. Finally, we denote by $\mathcal{N} := S_1^{Tot}$ the maximum number of accessible states in 1-step (neighbors) from any state.

Regret and Uniform-PAC. We consider an agent that repeatedly interacts with an MDP in a sequence of episodes $[K]$. We denote by s_t^k and a_t^k, the state and action taken at the time step t of the k'th episode. We denote by $\mathcal{F}_{k-1}$, the filtration that includes all the events (states, actions, and rewards) until the end of the $(k-1)$'th episode, as well as the initial state of the k'th episode. Throughout the paper, we denote by π_k the policy that is executed during the k'th episode and assume it is $\mathcal{F}_{k-1}$ measurable. The performance of an agent is measured by its *regret*, defined as $\mathrm{Regret}(K) := \sum_{k=1}^K \big(V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)\big)$, as well as by the *Uniform-PAC* criterion [10], which we generalize to deal with approximate convergence. Let $\epsilon, \delta > 0$ and $N_\epsilon = \sum_{k=1}^\infty \mathbb{1}\big\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \epsilon\big\}$ be the number of episodes in which the algorithm outputs a policy whose value is ϵ-inferior to the optimal value. An algorithm is called Uniform-PAC, if $\Pr(\exists\epsilon > 0 : N_\epsilon \geq F(S, 1/\epsilon, \log 1/\delta, H)) \leq \delta$, where $F(\cdot)$ depends polynomially (at most) on its parameters. Note that Uniform-PAC implies (ϵ,δ)-PAC, and thus, it is a stronger property. As we analyze algorithms with inherent errors in this paper, we use a more general notion of Δ-Uniform-PAC by defining the random variable $N_\epsilon^\Delta = \sum_{k=1}^\infty \mathbb{1}\big\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \Delta + \epsilon\big\}$, where $\Delta > 0$. Finally, we use $\tilde{\mathcal{O}}(x)$ to represent x up to constants and poly-logarithmic factors in δ, and $O(x)$ to represent x up to constants.

3 Computing h-Lookahead Policies

Computing an action returned by a h-lookahead policy at a certain state is a main component in the RTDP-based algorithms we analyze in Sections 5 and 6. A 'naive' procedure that returns such action is the exhaustive search. Its computational cost is $O(A^h)$ and $O(A^{Sh})$ for deterministic and stochastic systems, respectively. Such an approach is impractical, even for moderate values of h or S.

Instead of the naive approach, we formulate a Forward-Backward DP (FB-DP) algorithm, whose pseudo-code is given in Appendix 9. The FB-DP returns an action of an h-lookahead policy from a given state s. Importantly, in both deterministic and stochastic systems, the computation cost of FB-DP depends linearly on the total *number of reachable states* from s in up to h steps, i.e., $S_h^{Tot}(s)$. In the worst case, we may have $S_h(s) = O(\min(A^h, S))$. However, when $S_h^{Tot}(s)$ is small, significant improvement is achieved by avoiding unnecessary repeated computations.

FB-DP has two subroutines. It first constructs the set of reachable states from state s in up to h steps, $\{\mathcal{S}_t(s)\}_{t=1}^h$, in the 'forward-pass'. Given this set, in the second 'backward-pass' it simply applies backward induction (Eq. 3) and returns an action suggested by the h-lookahead policy, $a_h(s)$. Note that at each stage $t \in [h]$ of the backward induction (applied on the set $\{\mathcal{S}_t(s)\}_{t=1}^h$) there are $S_t(s)$ states on which the Bellman operator is applied. Since applying the Bellman operator costs $O(\mathcal{N}A)$ computations, the computational cost of the 'backward-pass' is $O\big(\mathcal{N}AS_h^{Tot}(s)\big)$.

In Appendix 9, we describe a DP-based approach to efficiently implement 'forward-pass' and analyze its complexity. Specifically, we show the computational cost of the 'forward-pass' is equivalent to that of the 'backward-pass' (see Propsition 8). Meaning, the computational cost of FB-DP is $O\big(\mathcal{N}AS_h^{Tot}(s)\big)$ - same order as the cost of backward induction given the set $\mathcal{S}_h^{Tot}(s)$.

4 Real-Time Dynamic Programming

Real-time dynamic programming (RTDP) [3] is a well-known online planning algorithm that assumes access to a transition model and a reward function. Unlike DP algorithms (policy, value iteration, or asynchronous value iteration) [4] that solve an MDP using offline calculations and sweeps over the entire states (possibly in random order), RTDP solves it in real-time, using samples from the environment (either simulated or real) and DP-style Bellman updates from the current state. Furthermore, unlike DP algorithms, RTDP needs to tradeoff exploration-exploitaion, since it interacts with the environment via sampling trajectories. This makes RTDP a good candidate for problems in which having access to the entire state space is not possible, but interaction is.

Algorithm 1 contains the pseudo-code of RTDP in finite-horizon MDPs. The value is initialized optimistically, $\bar{V}^0_{t+1}(s) = H - t \geq V^*_{t+1}(s)$. At each time step $t \in [H]$ and episode $k \in [K]$, the agent updates the value of the current state s^k_t by the optimal Bellman operator. It then acts greedily w.r.t. the current value at the next time step $\bar{V}^{k-1}_{t+1}$. Finally, the next state, s^k_{t+1}, is sampled either from the model or the real-world. When the model is exact, there is no difference in sampling from the model and real-world, but these are different in case the model is inexact as in Section 6.1.

The following high probability bound on the regret of a Decreasing Bounded Process (DBP), proved in [15], plays a key role in our analysis of exact and approximate RTDP with lookahead policies in Sections 5 and 6. An adapted process $\{X_k, \mathcal{F}_k\}_{k\geq 0}$ is a DBP, if for all $k \geq 0$, **(i)** $X_k \leq X_{k-1}$ almost surely (a.s.), **(ii)** $X_k \geq C_2$, and **(iii)** $X_0 = C_1 \geq C_2$. Interestingly, contrary to the standard regret bounds (e.g., in bandits), this bound does not depend on the number of rounds K.

Theorem 1 (Regret Bound of a DBP [15])**.** *Let $\{X_k, \mathcal{F}_k\}_{k\geq 0}$ be a DBP and $R_K = \sum_{k=1}^K X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]$ be its K-round regret. Then,*

$$\Pr\{\exists K > 0 : R_K \geq 9(C_1 - C_2)\ln(3/\delta)\} \leq \delta.$$

5 RTDP with Lookahead Policies

In this section, we devise and analyze a lookahead-based RTDP algorithm, called h-RTDP, whose pseudo-code is shown in Algorithm 2. Without loss of generality, we assume that $H/h \in \mathbb{N}$. We divide the horizon H into H/h intervals, each of length h time steps. h-RTDP stores HS/h values in the memory, i.e., the values at time steps $\mathcal{H} = \{1, h+1, \ldots, H+1\}$.[7] For each time step $t \in [H]$, we denote by $h_c \in \mathcal{H}$, the next time step for which a value is stored in the memory, and by $t_c = h_c - t$, the number of time steps until there (see Figure 1). At each time step t of an episode $k \in [K]$, given the current state s^k_t, h-RTDP selects an action a^k_t returned by the t_c-lookahead policy w.r.t. $\bar{V}^{k-1}_{h_c}$,

$$a^k_t = a_{t_c}(s^k_t) \in \arg\max_{\pi_1(s^k_t)} \max_{\pi_2,\ldots,\pi_{t_c}} \mathbb{E}\Big[\sum_{t'=1}^{t_c} r(s_{t'}, \pi_{t'}(s_{t'})) + \bar{V}^{k-1}_{h_c}(s_{t_c+1}) \mid s_1 = s^k_t\Big]. \qquad (4)$$

Thus, h-RTDP uses a varying lookahead horizon t_c that depends on how far the current time step is to the next one for which a value is stored. Throughout the paper, with an abuse of notation, we refer to this policy as a h-lookahead policy. Finally, it can be seen that h-RTDP generalizes RTDP as they are equal for $h = 1$.

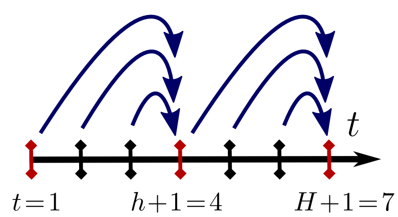

Figure 1: Varying lookahead horizon of a h-greedy policy in h-RTDP (see Eq. 4) with $h = 3$ and $H = 6$. The blue arrows show the lookahead horizon from a specific time step t, and the red bars are the time steps for which a value is stored in memory, i.e., $\mathcal{H} = \{1\ ,\ h+1 = 4\ ,\ 2h+1 = H+1 = 7\}$.

We are now ready to establish finite-sample performance guarantees for h-RTDP; see Appendix 10 for the detailed proofs. We start with two lemmas from which we derive the main convergence result of this section.

Lemma 2. *For all $s \in \mathcal{S}$, $n \in \{0\} \cup [\frac{H}{h}]$, and $k \in [K]$, the value function of h-RTDP is (i) Optimistic: $V^*_{nh+1}(s) \leq \bar{V}^k_{nh+1}(s)$, and (ii) Non-Increasing: $\bar{V}^k_{nh+1}(s) \leq \bar{V}^{k-1}_{nh+1}(s)$.*

[7] In fact, h-RTDP does not need to store V_1 and V_{H+1}, they are only used in the analysis.

Algorithm 1 Real-Time DP (RTDP)

```
init: ∀s ∈ S, ∀t ∈ {0} ∪ [H],
      V̄^0_{t+1}(s) = H − t
for k ∈ [K] do
  Initialize s^k_1 arbitrarily
  for t ∈ [H] do
    V̄^k_t(s^k_t) = T V̄^{k−1}_{t+1}(s^k_t)
    a^k_t ∈ arg max_a r(s^k_t, a) + p(·|s^k_t, a) V̄^{k−1}_{t+1}
    Act by a^k_t, observe s^k_{t+1} ∼ p(· | s^k_t, a^k_t)
  end for
end for
```

Algorithm 2 RTDP with Lookahead (h-RTDP)

```
init:: ∀s ∈ S, n ∈ {0} ∪ [H/h],
       V̄^0_{nh+1}(s) = H − nh
for k ∈ [K] do
  Initialize s^k_1 arbitrarily
  for t ∈ [H] do
    if (t − 1) mod h = 0 then
      h_c = t + h
      V̄^k_t(s^k_t) = T^h V̄^{k−1}_{h_c}(s^k_t)
    end if
    a^k_t ∈
    arg max_a r(s^k_t, a) + p(·|s^k_t, a) T^{h_c−t−1} V̄^{k−1}_{h_c}
    Act by a^k_t, observe s^k_{t+1} ∼ p(· | s^k_t, a^k_t)
  end for
end for
```

Lemma 3 (Optimality Gap and Expected Decrease)**.** *The expected cumulative value update at the k'th episode of h-RTDP satisfies* $\bar{V}^k_1(s^k_1) - V^{\pi_k}_1(s^k_1) = \sum_{n=1}^{\frac{H}{h}-1} \sum_{s\in\mathcal{S}} \bar{V}^{k-1}_{nh+1}(s) - \mathbb{E}[\bar{V}^k_{nh+1}(s) \mid \mathcal{F}_{k-1}]$.

Properties (i) and (ii) in Lemma 2 show that $\{\bar{V}^k_{nh+1}(s)\}_{k\geq 0}$ is a DBP, for any s and n. Lemma 3 relates $\bar{V}^k_1(s^k_1) - V^{\pi_k}_1(s^k_1)$ (LHS) to the expected decrease in $\bar{V}^k$ at the k'th episode (RHS). When the LHS is small, then $\bar{V}^k_1(s^k_1) \simeq V^*_1(s^k_1)$, due to the optimism of $\bar{V}^k_1$, and h-RTDP is about to converge to the optimal value. This is why we refer to the LHS as the *optimality gap*. Using these two lemmas and the regret bound of a DBP (Theorem 1), we prove a finite-sample convergence result for h-RTDP (see Appendix 10 for the full proof).

Theorem 4 (Performance of h-RTDP)**.** *Let $\epsilon, \delta > 0$. The following holds for h-RTDP:*

1. *With probability $1-\delta$, for all $K > 0$,* $\mathrm{Regret}(K) \leq \frac{9SH(H-h)}{h}\ln(3/\delta)$.
2. $\Pr\Big\{\exists \epsilon > 0 \ :\ N_\epsilon \geq \frac{9SH(H-h)\ln(3/\delta)}{h\epsilon}\Big\} \leq \delta$.

Proof Sketch. Applying Lemmas 2 and 3, we may write

$$\begin{aligned}\mathrm{Regret}(K) &\leq \sum_{k=1}^{K} \bar{V}^k_1(s^k_1) - V^{\pi_k}_1(s^k_1) = \sum_{k=1}^{K} \sum_{n=1}^{\frac{H}{h}-1} \sum_{s} \bar{V}^{k-1}_{nh+1}(s) - \mathbb{E}[\bar{V}^k_{nh+1}(s) \mid \mathcal{F}_{k-1}] \\ &= \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]. \end{aligned} \tag{5}$$

Where we define $X_k := \sum_{n=1}^{\frac{H}{h}-1} \sum_s \bar{V}^{k-1}_{nh+1}(s)$ and use linearity of expectation. By Lemma 2, $\{X_k\}_{k\geq 0}$ is decreasing and bounded from below by $\sum_{n=1}^{\frac{H}{h}-1} \sum_s V^*_{nh+1}(s) \geq 0$. We conclude the proof by observing that $X_0 \leq \sum_{n=1}^{\frac{H}{h}-1} \sum_s V^0_{nh+1}(s) \leq SH(H-h)/h$, and applying Theorem 1. □

Remark 1 (RTDP and Good Value Initialization)**.** *A closer look into the proof of Theorem 4 shows we can easily obtain a stronger result which depends on the initial value V^0. The regret can be bounded by* $\mathrm{Regret}(K) \leq \tilde{\mathcal{O}}\Big(\sum_{n=1}^{\frac{H}{h}-1}\big(V^0_{nh+1}(s) - V^*_{nh+1}(s)\big)\Big)$, *which formalizes the intuition the algorithm improves as the initial value V^0 better estimates V^*. For clarity purposes we provide the worse-case bound.*

Remark 2 (Computational Complexity of h-RTDP)**.** *Using FB-DP (Section 3) as a solver of a h-lookahead policy, the per-episode* computation cost *of h-RTDP amounts to applying FB-DP for H time steps, i.e., it is bounded by $O(H\mathcal{N}AS^{Tot}_h)$. Since S^{Tot}_h – the total number of reachable states in up to h time steps – is an increasing function of h, the computation cost of h-RTDP increases with h, as expected. When S^{Tot}_h is significantly smaller than S, the per-episode computational complexity of*

h-RTDP is S independent. As discussed in Section 3, using FB-DP, in place of exhaustive search, can significantly improve the computational cost of h-RTDP.

Remark 3 (Improved Sample Complexity of h-RTDP)**.** *Theorem 4 shows that h-RTDP improves the* sample complexity *of RTDP by a factor $1/h$. This is consistent with the intuition that larger horizon of the applied lookahead policy results in faster convergence (less samples). Thus, if RTDP is used in a real-time manner, one way to boost its performance is to combine with lookahead policies.*

Remark 4 (Sparse Sampling Approaches)**.** *In this work, we assume h-RTDP has access to a h-lookahead policy* (3) *solver, such as FB-DP presented in Section 3. We leave studying the sparse sampling approach [19, 28] for approximately solving h-lookahead policy for future work.*

6 Approximate RTDP with Lookahead Policies

In this section, we consider three approximate versions of h-RTDP in which the update deviates from its exact form described in Section 5. We consider the cases in which there are errors in the **1)** *model*, **2)** *value updates*, and when we use **3)** *approximate state abstraction*. We prove finite-sample bounds on the performance of h-RTDP in the presence of these approximations. Furthermore, in Section 6.3, given access to an approximate state abstraction, we show that the convergence of h-RTDP depends on the cardinality of the *abstract state space* – which can be much smaller than the original one. The proofs of this section generalize that of Theorem 4, while following the same 'recipe'. This shows the generality of the proof technique, as it works for both exact and approximate settings.

6.1 h-RTDP with Approximate Model (h-RTDP-AM)

In this section, we analyze a more practical scenario in which the transition model used by h-RTDP to act and update the values is not exact. We assume it is close to the true model in the total variation (TV) norm, $\forall (s,a) \in \mathcal{S}\times\mathcal{A},\ ||p(\cdot|s,a) - \hat{p}(\cdot|s,a)||_1 \leq \epsilon_P$, where $\hat{p}$ denotes the approximate model. Throughout this section and the relevant appendix (Appendix 11), we denote by $\hat{T}$ and $\hat{V}^*$ the optimal Bellman operator and optimal value of the approximate model $\hat{p}$, respectively. Note that $\hat{T}$ and $\hat{V}^*$ satisfy (1) and (2) with p replaced by $\hat{p}$. h-RTDP-AM is exactly the same as h-RTDP (Algorithm 2) with the model p and optimal Bellman operator T replaced by their approximations $\hat{p}$ and $\hat{T}$. We report the pseudocode of h-RTDP-AM in Appendix 11.

Although we are given an approximate model, $\hat{p}$, we are still interested in the performance of (approximate) h-RTDP on the *true MDP*, p, and relative to its optimal value, V^*. If we solve the approximate model and act by its optimal policy, the Simulation Lemma [20, 30] suggests that the regret is bounded by $O(H^2\epsilon_P K)$. For h-RTDP-AM, the situation is more involved, as its updates are based on the approximate model and the samples are gathered by interacting with the true MDP. Nevertheless, by properly adjusting the techniques from Section 5, we derive performance bounds for h-RTDP-AM. These bounds reveal that the asymptotic regret increases by at most $O(H^2\epsilon_P K)$, similarly to the regret of the optimal policy of the approximate model. Interestingly, the proof technique follows that of the exact case in Theorem 4. We generalize Lemmas 2 and 3 from Section 5 to the case that the update rule uses an inexact model (see Lemmas 9 and 10 in Appendix 11). This allows us to establish the following performance bound for h-RTDP-AM (proof in Appendix 11).

Theorem 5 (Performance of h-RTDP-AM)**.** *Let $\epsilon, \delta > 0$. The following holds for h-RTDP-AM:*

1. *With probability $1-\delta$, for all $K > 0$,* $\mathrm{Regret}(K) \leq \frac{9SH(H-h)}{h}\ln(3/\delta) + H(H-1)\epsilon_P K.$
2. *Let $\Delta_P = H(H-1)\epsilon_P$. Then,* $\Pr\left\{\exists \epsilon > 0 \ :\ N_\epsilon^{\Delta_P} \geq \frac{9SH(H-h)\ln(3/\delta)}{h\epsilon}\right\} \leq \delta.$

These bounds show the approximate convergence resulted from the approximate model. However, the asymptotic performance gaps – both in terms of the regret and Uniform PAC – of h-RTDP-AM approach those experienced by an optimal policy of the approximate model. Interestingly, although h-RTDP-AM updates using the approximate model, while interacting with the true MDP, its convergence rate (to the asymptotic performance) is similar to that of h-RTDP (Theorem 4).

6.2 h-RTDP with Approximate Value Updates (h-RTDP-AV)

Another important question in the analysis of approximate DP algorithms is their performance under approximate value updates, motivated by the need to use function approximation. This is often modeled by an extra noise $|\epsilon_V(s)| \leq \epsilon_V$ added to the update rule [4]. Following this approach, we study such perturbation in h-RTDP. Specifically, in h-RTDP-AV the value update rule is modified such that it contains an error term (see Algorithm 2),

$$\bar{V}_t^k(s_t^k) = \epsilon_V(s_t^k) + T^h \bar{V}_{h_c}^{k-1}(s_t^k).$$

For $\epsilon_V(s_t^k) = 0$, the exact h is recovered. The pseudocode of h-RTDP-AV is supplied in Appendix 12.

Similar to the previous section, we follow the same proof technique as for Theorem 4 to establish the following performance bound for h-RTDP-AV (proof in Appendix 12).

Theorem 6 (Performance of h-RTDP-AV)**.** *Let $\epsilon, \delta > 0$. The following holds for h-RTDP-AV:*

1. *With probability $1-\delta$, for all $K > 0$,* $\text{Regret}(K) \le \frac{9SH(H-h)}{h}(1 + \frac{H}{h}\epsilon_V)\ln(\frac{3}{\delta}) + \frac{2H}{h}\epsilon_V K.$
2. *Let $\Delta_V = 2H\epsilon_V$. Then,* $\Pr\left\{\exists \epsilon > 0 \ :\ N_\epsilon^{\frac{\Delta_V}{h}} \ge \frac{9SH(H-h)(1+\frac{\Delta_V}{2h})\ln(\frac{3}{\delta})}{h\epsilon}\right\} \le \delta.$

As in Section 6.1, the results of Theorem 6 exhibit an asymptotic linear regret $O(H\epsilon_V K/h)$. As proven in Proposition 20 in Appendix 15, such performance gap exists in ADP with approximate value updates. Furthermore, the convergence rate in S to the asymptotic performance of h-RTDP-AV is similar to that of its exact version (Theorem 4). Unlike in h-RTDP-AM, the asymptotic performance of h-RTDP-AV *improves* with h. This quantifies a clear benefit of using lookahead policies in online planning when the value function is approximate.

6.3 h-RTDP with Approximate State Abstraction (h-RTDP-AA)

We conclude the analysis of approximate h-RTDP with exploring the advantages of combining it with approximate state abstraction [1]. The central result of this section establishes that given an approximate state abstraction, h-RTDP converges with sample, computation, and space complexity *independent* of the size of the state space S, as long as S_h^{Tot} is smaller than S (i.e., when performing h-lookahead is S independent, Remark 2). This is in contrast to the computational complexity of ADP in this setting, which is still $O(HSA)$ (see Appendix 15.3 for further discussion). State abstraction has been widely investigated in approximate planning [12, 11, 16, 1], as a means to deal with large state space problems. Among existing approximate abstraction settings, we focus on the following one. For any $n \in \{0\} \cup [\frac{H}{h} - 1]$, we define $\phi_{nh+1} : \mathcal{S} \to \mathcal{S}_\phi$ to be a mapping from the state space $\mathcal{S}$ to reduced space $\mathcal{S}_\phi$, $S_\phi = |\mathcal{S}_\phi| \ll S$. We make the following assumption:

Assumption 1 (Approximate Abstraction, [23], definition 3.3)**.** *For any $s, s' \in \mathcal{S}$ and $n \in \{0\} \cup [\frac{H}{h} - 1]$ for which $\phi_{nh+1}(s) = \phi_{nh+1}(s')$, we have $|V_{nh+1}^*(s) - V_{nh+1}^*(s')| \le \epsilon_A$.*

Let us denote by $\{\bar{V}_{\phi,nh+1}^k\}_{n=0}^{H/h}$ the values stored in memory by h-RTDP-AA at the k'th episode. Unlike previous sections, the value function per time step contains S_ϕ entries, $\bar{V}_{\phi,1+nh}^k \in \mathbb{R}^{S_\phi}$. Note that if $\epsilon_A = 0$, then optimal value function can be represented in the reduced state space $\mathcal{S}_\phi$. However, if ϵ_A is positive, exact representation of V^* is not possible. Nevertheless, the asymptotic performance of h-RTDP-AA will be 'close', up to error of ϵ_A, to the optimal policy.

Furthermore, the definition of the multi-step Bellman operator (2) and h-greedy policy (3) should be revised, and with some abuse of notation, defined as

$$a_t^k \in \arg\max_{\pi_0(s_t^k)} \max_{\pi_1,\ldots,\pi_{t_c-1}} \mathbb{E}\left[\sum_{t'=0}^{t_c-1} r_{t'} + \bar{V}_{\phi,h_c}^{k-1}(\phi_{h_c}(s_{t_c})) \mid s_0 = s_t^k\right], \tag{6}$$

$$T_\phi^h \bar{V}_{\phi,h_c}^{k-1}(s_t^k) := \max_{\pi_0,\ldots,\pi_{h-1}} \mathbb{E}\left[\sum_{t'=0}^{h-1} r_{t'} + \bar{V}_{\phi,t+h}^{k-1}(\phi_{t+h}(s_h)) \mid s_0 = s_t^k\right]. \tag{7}$$

Eq. (6) and (7) indicate that similar to (3), the h-lookahead policy uses the given model to plan for h time steps ahead. Differently from (3), the value after h time steps is the one defined in the *reduced state* space $\mathcal{S}_\phi$. Note that the definition of the h-greedy policy for h-RTDP-AA in (6) is equivalent to the one used in Algorithm 8, obtained by similar recursion as for the optimal Bellman operator (2).

h-RTDP-AA modifies both the value update and the calculation of the h-lookahead policy (the value update and action choice in algorithm 2). The h-lookahead policy is replaced by h-lookahead defined in (6). The value update is substituted by (7), i.e, $\bar{V}_{\phi,t}^k(\phi_t(s_t^k)) = T_\phi^h \bar{V}_{\phi,h_c}^{k-1}(s_t^k)$. The full pseudocode of h-RTDP-AA is supplied in Appendix 13. By similar technique, as in the proof of Theorem 4, we establish the following performance guarantees to h-RTDP-AA (proof in Appendix 13).

Setting	h-RTDP Regret (This work)	ADP Regret [4]	UCT
Exact (5)	$\tilde{\mathcal{O}}\big(SH(H-h)/h\big)$	0	$\Omega(\exp(\exp(H)))$ [9]
App. Model (6.1)	$\tilde{\mathcal{O}}\big(SH(H-h)/h+\Delta_P K\big)$	$\Delta_P K$	N.A
App. Value (6.2)	$\tilde{\mathcal{O}}\big(SH(H-h)g^{\epsilon}_{H/h}/h+\Delta_V K/h\big)$	$\Delta_V K/h$	N.A
App. Abstraction (6.3)	$\tilde{\mathcal{O}}\big(S_\phi H(H-h)/h+\Delta_A K/h\big)$	$\Delta_A K/h$	N.A

Table 1: The lookhead horizon is h and the horizon of the MDP is H. We denote $g^{\epsilon}_{H/h} = (1 + H\epsilon_V/h)$, $\Delta_P = H(H-1)\epsilon_P$, $\Delta_V = 2H\epsilon_V$, and $\Delta_A = H\epsilon_A$. The table summarizes the regret bounds of the h-RTDP settings studied in this work and compares them to those of their corresponding ADP approaches. The performance of ADP is based on standard analysis, supplied in Propositions 19, 20, 21 in Appendix 15.

Theorem 7 (Performance of h-RTDP-AA)**.** *Let $\epsilon, \delta > 0$. The following holds for h-RTDP-AA:*

1. *With probability* $1-\delta$, *for all* $K > 0$, $\quad \mathrm{Regret}(K) \leq \frac{9S_\phi H(H-h)}{h}\ln(3/\delta) + \frac{H\epsilon_A}{h}K.$
2. *Let $\Delta_A = H\epsilon_A$. Then,* $\quad \Pr\left\{\exists \epsilon > 0 \;:\; N_\epsilon^{\frac{\Delta_A}{h}} \geq \frac{9S_\phi H(H-h)\ln(3/\delta)}{h\epsilon}\right\} \leq \delta.$

Theorem 7 establishes S-independent performance bounds that depend on the size of the reduced state space S_ϕ. The asymptotic regret and Uniform PAC guarantees are approximate, as the state abstraction is approximate. Furthermore, they are improving with the quality of approximation ϵ_A, i.e., their asymptotic gap is $O(H\epsilon_A/h)$ relative to the optimal policy. Moreover, the asymptotic performance of h-RTDP-AA improves as h is increased. Importantly, since the computation complexity of each episode of h-RTDP is independent of S (Section 3), the computation required to reach the approximate solution in h-RTDP-AA is also S-independent. This is in contrast to the computational cost of DP that depends on S and is $O(SHA)$ (see Appendix 15.3 for further discussion).

7 Discussion and Conclusions

RTDP vs. DP. The results of Sections 5 and 6 established finite-time convergence guarantees for the exact h-RTDP and its three approximations. In the approximate settings, as expected, the regret has a linear term of the form ΔK, where Δ is linear in the approximation errors ϵ_P, δ, and ϵ_A, and thus, the performance is continuous in these parameters, as we would desire. We refer to ΔK as the *asymptotic regret*, since it dominates the regret as $K \to \infty$.

A natural measure to evaluate the quality of h-RTDP in the approximate settings is comparing its regret to that of its corresponding approximate DP (ADP). Table 1 summarizes the regrets of the approximate h-RTDPs studied in this paper and their corresponding ADPs. ADP calculates approximate values $\{V^*_{nh+1}\}_{n=0}^{H/h}$ by backward induction. Based on these values, the same h-lookahead policy by which h-RTDP acts is evaluated. In the analysis of ADP, we use standard techniques developed for the discounted case in [4]. From Table 1, we reach the following conclusion: *the asymptotic performance (in terms of regret) of approximate h-RTDP is equivalent to that of a corresponding approximate DP algorithm*. Furthermore, it is important to note that the asymptotic error decreases with h for the approximate value updates and approximate abstraction settings for both RTDP and DP algorithms. In these settings, the error is caused by approximation in the value function. By increasing the lookahead horizon h, the algorithm uses less such values and relies more on the model which is assumed to be correct. Thus, the algorithm becomes less affected by the value function approximation.

Conclusions. In this paper, we formulated h-RTDP, a generalization of RTDP that acts by a lookahead policy, instead of by a 1-step greedy policy, as in RTDP. We analyzed the finite-sample performance of h-RTDP in its exact form, as well as in three approximate settings. The results indicate that h-RTDP converges in a very strong sense. Its regret is constant w.r.t. to the number of episodes, unlike in, e.g., reinforcement learning where a lower bound of $\tilde{\mathcal{O}}(\sqrt{SAHT})$ exists [2, 18]. Furthermore, the analysis reveals that the sample complexity of h-RTDP improves by increasing the lookahead horizon h (Remark 3). Moreover, the asymptotic performance of h-RTDP was shown to be equivalent to that of ADP (Table 1), which under no further assumption on the approximation error, is the best we can hope for.

We believe this work opens interesting research venues, such as studying alternatives to the solution of the h-greedy policy (see Section 9), studying a Receding-Horizon extension of RTDP, RTDP with function approximation, and formulating a Thompson-Sampling version of RTDP, as the standard RTDP is an 'optimistic' algorithm. As the analysis developed in this work was shown to be quite generic, we hope that it can assist with answering some of these questions. On the experimental side, more needs to be understood, especially comparing RTDP with MCTS and studying how RTDP can be combined with deep neural networks as the value function approximator.

8 Broader Impact

Online planning algorithms, such as A^* and RTDP, have been extensively studied and applied in AI for well over two decades. Our work quantifies the benefits of using lookahead-policies in this class of algorithms. Although lookahead-policies have also been widely used in online planning algorithms, their theoretical justification was lacking. Our study sheds light on the benefits of lookahead-policies. Moreover, the results we provide in this paper suggest improved ways for applying lookahead-policies in online planning with benefits when dealing with various types of approximations. This work opens up the room for practitioners to improve their algorithms and base lookahead policies on solid theoretical ground.

Acknowledgements. We thank the reviewers for their helpful comments and feedback. Y.E. and S.M. were partially supported by the ISF under contract 2199/20.

References

[1] David Abel, D. Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, pages 2915–2923, 2016.

[2] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

[3] Andrew Barto, Steven Bradtke, and Satinder Singh. Learning to act using real-time dynamic programming. *Artificial intelligence*, 72(1-2):81–138, 1995.

[4] D. Bertsekas and J. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[5] Blai Bonet and Hector Geffner. Planning with incomplete information as heuristic search in belief space. In *Proceedings of the Fifth International Conference on Artificial Intelligence Planning Systems*, pages 52–61. AAAI Press, 2000.

[6] Blai Bonet and Hector Geffner. Labeled rtdp: Improving the convergence of real-time dynamic programming. In *ICAPS*, volume 3, pages 12–21, 2003.

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

[8] Vadim Bulitko and Greg Lee. Learning in real-time search: A unifying framework. *Journal of Artificial Intelligence Research*, 25:119–157, 2006.

[9] Pierre-Arnaud Coquelin and Rémi Munos. Bandit algorithms for tree search. *arXiv preprint cs/0703062*, 2007.

[10] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

[11] Thomas Dean, Robert Givan, and Sonia Leach. Model reduction techniques for computing approximately optimal solutions for Markov decision processes. In *Proceedings of the 13th conference on Uncertainty in artificial intelligence*, pages 124–131, 1997.

[12] Richard Dearden and Craig Boutilier. Abstraction and approximate decision-theoretic planning. *Artificial Intelligence*, 89(1-2):219–283, 1997.

[13] Y. Efroni, G. Dalal, B. Scherrer, and S. Mannor. How to combine tree-search methods in reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3494–3501, 2019.

[14] Yonathan Efroni, Gal Dalal, Bruno Scherrer, and Shie Mannor. Multiple-step greedy policies in approximate and online reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5238–5247, 2018.

[15] Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pages 12203–12213, 2019.

[16] Eyal Even-Dar and Yishay Mansour. Approximate equivalence of markov decision processes. In *Learning Theory and Kernel Machines*, pages 581–594, 2003.

[17] Matthieu Geist and Olivier Pietquin. Algorithmic survey of parametric value function approximation. *IEEE Transactions on Neural Networks and Learning Systems*, 24(6):845–867, 2013.

[18] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

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

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

[21] Levente Kocsis and Csaba Szepesvári. Bandit based Monte-Carlo planning. In *European conference on machine learning*, pages 282–293, 2006.

[22] Andrey Kolobov, Daniel S Weld, et al. Lrtdp versus uct for online probabilistic planning. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[23] L. Li, T. Walsh, and M. Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pages 531–539, 2006.

[24] Brendan McMahan, Maxim Likhachev, and Geoffrey Gordon. Bounded real-time dynamic programming: Rtdp with monotone upper bounds and performance guarantees. In *Proceedings of the 22nd international conference on Machine learning*, pages 569–576. ACM, 2005.

[25] Rémi Munos. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

[26] Rémi Munos. From bandits to Monte-Carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014.

[27] Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, and Matthieu Geist. Approximate modified policy iteration. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1207–1214, 2012.

[28] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787, 2018.

[29] A. Strehl, L. Li, and M. Littman. PAC reinforcement learning bounds for RTDP and rand-RTDP. In *Proceedings of AAAI workshop on learning for search*, 2006.

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

9 Per-Episode Complexity of h-RTDP

In this section, we define and analyze the *Forward-Backward DP* by which an h-greedy policy can be calculated from a current state s_t^k according to (3). Observe that the algorithm is based on a 'local' information, i.e., it does not need access to the entire state space, but to a portion of the state space in the 'vicinity' of the current state s_t^k. Furthermore, it does not assume prior knowledge on this vicinity.

9.1 Forward-Backward Dynamic Programming Approach

Algorithm 3 h-Forward-Backward DP

```
Input: s, transition p, reward r, lookahead horizon h, value at the end of lookahead horizon V̄
{S_t'(s)}_{t'=1}^{h+1} = Forward-Pass(s,p, h)
action = Backward-Pass({S_t'(s)}_{t'=1}^{h+1}, r, p, h, V̄)
return: action
```

Algorithm 4 Forward-Pass

```
Input: Starting state s, p, h
Init: S_1 = {s}, ∀t' ∈ [h]/{1}, S_t'(s) = {}
for t' = 2, 3, ..., h+1 do
   for s_{t'-1} ∈ S_{t'-1}(s) do
      # acquire possible next states from s_{t'-1}
      for a ∈ A do
         S_t'(s) = S_t'(s) ∪ {s' : p(s' | s, a) > 0}
      end for
   end for
end for
return: {S_t'(s)}_{t'=1}^{h+1}
```

Algorithm 5 Backward-Pass

```
Input: {S_t'(s)}_{t'=1}^{h+1}, r, p, h, V̄
# initialize values by arbitrary value C
Init: ∀t' ∈ [h-1], ∀s ∈ S_t'(s), V_t'(s) = C
# Assign the value at t' = h to the current value, V.
for s ∈ S_{h+1}(s) do
   V_{h+1}(s) = V̄(s)
end for
for t' = h, h-1, ..., 2 do
   for s ∈ S_t'(s) do
      V_t'(s) = max_a r(s,a) + p(· | s,a) V_{t'+1}
   end for
end for
return: arg max_a r(s,a) + p(· | s,a) V_2
```

The Forward-Backword DP (Algorithm 3) approach is built on the following observation: would we known the accessible state space from s in next h time steps we could use Backward Induction (i.e., Value Iteration) on a finite-horizon MDP, with an horizon of h, and calculate the optimal policy from s. Unfortunately, as we do not assume such a prior knowledge, we have to calculate this set before applying the backward induction step. Thus, Forward-Backword DP first build this set (in the first, 'Forward' stage) and later applies standard backward induction (in the 'Backward' stage). In Proposition 8, we establish that calculating the set of accessible states can be done efficiently

Let us first analyze the *computational complexity* of Algorithm 3 using the following definitions. Let $\mathcal{S}_{t'}(s)$ be the set of reachable states from state s in t' times steps, formally,

$$\mathcal{S}_{t'}(s) = \{s' \mid \exists \pi : p^\pi(s_{t'} = s' \mid s_0 = s, \pi) > 0\},$$

where $p^\pi(s_{t'} = s' \mid s_0 = s, \pi) = \mathbb{E}[\mathbb{1}\{s_{t'} = s'\} \mid s_0 = s, \pi]$. The cardinality of this set is denoted by $|\mathcal{S}_{t'}(s)|$. let $\mathcal{N} := \max_s |\mathcal{S}_2(s)|$ be the maximal number of accessible states in 1-step (maximal 'nearest neighbors' from any state). Furthermore, let the total reachable states in h time steps from state s be $S_h^{Tot}(s) = \sum_{t'=1}^{h} |\mathcal{S}_{t'}(s)|$. When $S_h^{Tot}(s)$ is small, as we establish in this section, local search up to an horizon of h can be done efficiently with the Forward-Backward DP, unlike the exhaustive search approach.

Based on the above definitions we analyze the computational complexity of Forward-Backward DP starting from the Forward-Pass stage.

Proposition 8 (Computation Cost of Forward-Pass)**.** *The Forward-Pass stage of FB-DP can be implemented with the computation cost of $O\big(\mathcal{N} A S_h^{Tot}(s)\big)$.*

Proof. Calculating the set $\{s' : p(s' \mid s,a) > 0\}$ cost is upper bounded by $O(\mathcal{N})$ as we need to enumerate at most all possible $O(\mathcal{N})$ next-states. We assume that $\mathcal{S}_{t'} = \mathcal{S}_{t'}(s) \cup \{s' : p(s' \mid s,a) > 0\}$ can be done by $O(\mathcal{N})$, e.g., when using a hash-table for saving $\mathcal{S}_{t'}$ in memory. As we need to repeat this operation A times, the complexity for each $t' \in \{2,3,.,,h+1\}$ is upper bounded by $O(\mathcal{N}A|S_{t'-1}(s)|)$. Summing over all t' we get that the computational complexity of the Forward pass is upper bounded by

$$O\left(\mathcal{N}A\sum_{t'=2}^{h+1}|S_{t'-1}(s)|\right) = O\big(\mathcal{N}A\big|S_h^{Tot}(s)\big|\big),$$

where the second equality holds by definition of total number of accessible states in h time steps. □

The computational complexity of the backward passage is the computational complexity of Backward Induction, which is the total number of states in which actions can be taken times the number of actions per state, i.e.,

$$O(A\mathcal{N}S_h^{Tot}(s)).,\tag{8}$$

where the origin of the factor $\mathcal{N}$ is due to the need to calculate the sum $\sum_{s'} p(s' \mid s,a)V(s')$ for each (s,a) pair, and, by definition, this sum contain at most $\mathcal{N}$ elements.

Using Proposition 8 and (8) we get that for every $t \in [H]$, the computational complexity of calculating an h-lookahead policy from a state s using the Forward-Backward DP is bounded by,

$$O((\mathcal{N}A + \mathcal{N}A)S_h^{Tot}(s)) = O(\mathcal{N}AS_h^{Tot}),$$

where the last relation holds by definition, $S_h^{Tot} = \max_s S_h^{Tot}(s)$.

Finally, the *space complexity* of Forward-Backward DP is the space required the save in memory the possible visited states in h time steps (their identity in the Forward-Pass and their values in the Backward-Pass). By definition it is at most $O(hS_h)$.

10 Real Time Dynamic Programming with Lookahead

This section contains the full proofs of all the results of Section 5 in chronological order.

Lemma 2. *For all $s \in \mathcal{S}$, $n \in \{0\} \cup [\frac{H}{h}]$, and $k \in [K]$, the value function of h-RTDP is (i) Optimistic: $V^*_{nh+1}(s) \le \bar{V}^k_{nh+1}(s)$, and (ii) Non-Increasing: $\bar{V}^k_{nh+1}(s) \le \bar{V}^{k-1}_{nh+1}(s)$.*

Proof. Both claims are proven using induction.

(i) Let $n \in \{0\} \cup [\frac{H}{h}]$. By the initialization, $\forall s, n,\ V^*_{nh+1}(s) \le V^0_{nh+1}(s)$. Assume the claim holds for the first $(k-1)$ episodes. Let s^k_t be the state of the algorithm at a time step t of the k'th episode at which a value update takes place, i.e., $t = nh+1$, for some $n \in \{0\} \cup [\frac{H}{h}]$. By the value update of Algorithm 2 and (2), we have

$$\bar{V}^k_t(s^k_t) = (T^h \bar{V}^{k-1}_{h_c})(s^k_t) = (T^h \bar{V}^{k-1}_{t+h})(s^k_t) \ge (T^h V^*_{t+h})(s^k_t) = V^*_t(s^k_t).$$

The inequality holds by the induction hypothesis and the monotonicity of T^h, a consequence of the monotonicity of T, the optimal Bellman operator [4]. The last equality holds by the fact that the recursion is satisfied by the optimal value function (2). Thus, the induction step is proven for the first claim.

(ii) Let $n \in \{0\} \cup [\frac{H}{h}]$ and $t = nh+1$ be a time step in which a value update takes place. To prove the base case, we use the optimistic initialization. Let s^1_t be the state of the algorithm in the t'th time step of the first episode. By the update rule, we have

$$\bar{V}^1_t(s^1_t) = (T^h \bar{V}^0_{t+h})(s^0_t) = \max_{a_0,\dots,a_{h-1}} \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}) + \bar{V}^0_{t+h}(s_h) \mid s_0 = s^0_t\right]$$
$$\overset{(a)}{\le} h + H - (t+h-1) = H - (t-1) \overset{(b)}{=} \bar{V}^0_t(s^1_t).$$

(a) holds since $r(s,a) \in [0,1]$ and by the optimistic initialization.
(b) observe that $H-(t-1)$ is the value of the optimistic initialization.

Assume that the claim holds for the first $(k-1)$ episodes. Let s^k_t be the state of the algorithm at a time step t of the k'th episode at which a value update takes place, i.e., $t = nh+1$, for some $n \in \{0\} \cup [\frac{H}{h}]$. By the value update rule of Algorithm 2, we have $\bar{V}^k_t(s^k_t) = (T^h \bar{V}^{k-1}_{h_c})(s^k_t) = (T^h \bar{V}^{k-1}_{t+h})(s^k_t)$. If s^k_t was previously updated, let $\bar{k}$ be the last episode in which the update occurred, i.e., $\bar{V}^{\bar{k}}_t(s^k_t) = (T^h \bar{V}^{\bar{k}-1}_{t+h})(s^k_t) = \bar{V}^{k-1}_t(s^k_t)$. By the induction hypothesis, we have that $\forall s, t,\ \bar{V}^{\bar{k}-1}_t(s) \ge \bar{V}^{k-1}_t(s)$. Using the monotonicity of T^h, we may write

$$\bar{V}^k_t(s^k_t) = (T^h \bar{V}^{k-1}_{t+h})(s^k_t) \le (T^h \bar{V}^{\bar{k}-1}_{t+h})(s^k_t) = \bar{V}^{k-1}_t(s^k_t).$$

Thus, $\bar{V}^k_t(s^k_t) \le \bar{V}^{k-1}(s^k_t)$ and the induction step is proved. If s^k_t was not previously updated, then $\bar{V}^{k-1}_t(s^k_t) = \bar{V}^0_t(s^k_t)$. In this case, the induction hypothesis implies that $\forall s', \bar{V}^{k-1}_{t+h}(s') \le \bar{V}^0_{t+h}(s')$ and the result is proven similarly to the base case. □

Lemma 3 (Optimality Gap and Expected Decrease)**.** *The expected cumulative value update at the k'th episode of h-RTDP satisfies $\bar{V}^k_1(s^k_1) - V^{\pi_k}_1(s^k_1) = \sum_{n=1}^{\frac{H}{h}-1} \sum_{s \in \mathcal{S}} \bar{V}^{k-1}_{nh+1}(s) - \mathbb{E}[\bar{V}^k_{nh+1}(s) \mid \mathcal{F}_{k-1}]$.*

Proof. Let $n \in \{0\} \cup [\frac{H}{h}]$ and $t = nh+1$ be a time step in which a value update takes place. By the definition of the update rule, the following holds for the value update at the visited state s^k_t:

$$\bar{V}^k_t(s^k_t) = (T^h \bar{V}^{k-1}_{t+h})(s^k_t) \tag{9}$$
$$= (T^{\pi_k(t)} \cdots T^{\pi_k(t+h-1)} \bar{V}^{k-1}_{t+h})(s^k_t) = \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}, a_{t'}) + \bar{V}^{k-1}_{t+h}(s_{t+h}) \mid \pi_k, s_t = s^k_t\right]$$
$$\overset{(a)}{=} \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s^k_{t'}, a^k_{t'}) + \bar{V}^{k-1}_{t+h}(s^k_{t+h}) \mid \mathcal{F}_{k-1}, s^k_t\right]. \tag{10}$$

(a) We prove this passage for each reward element $r(s_{t'}, a_{t'})$ in the expectation. The proof for the expectation of $\bar{V}^{k-1}_{t+h}(s_{t+h})$ follows in a similar manner. Since the first expectation is w.r.t. the dynamics of the true model, a consequence of updating by the true model, for any $t' \geq t$, we may write

$$\mathbb{E}\big[r(s_{t'}, a_{t'}) \mid \pi_k, s_t = s^k_t\big] = \sum_{s_{t'} \in \mathcal{S}} p(s_{t'} \mid s_t = s^k_t, \pi_k) r(s_{t'}, \pi_k(s_{t'}, t'))$$
$$\overset{\text{(i)}}{=} \sum_{s^k_{t'} \in \mathcal{S}} p(s^k_{t'} \mid s^k_t, \mathcal{F}_{k-1}) r(s^k_{t'}, \pi_k(s^k_{t'}, t')) = \mathbb{E}\big[r(s^k_{t'}, a^k_{t'}) \mid \mathcal{F}_{k-1}, s^k_t\big],$$

where $p(s_{t'} \mid s^k_t, \pi_k)$ is the probability of starting at state s^k_t, following π_k, and reaching state $s_{t'}$ in $t' - t$ steps.

(i) We use the fact that $p(s_{t'} \mid s^k_t, \pi_k) = p(s^k_{t'} \mid s^k_t, \mathcal{F}_{k-1})$, in words, given the policy π_k (which is $\mathcal{F}_{k-1}$ measurable) and s^k_t the probability for a state $s^k_{t'}$ with $t' \geq t$ is independent of the rest of the history.

Now that we proved **(a)**, we take the conditional expectation of (9) w.r.t. $\mathcal{F}_{k-1}$ and use the tower rule to obtain

$$\mathbb{E}\big[\bar{V}^k_t(s^k_t) \mid \mathcal{F}_{k-1}\big] = \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s^k_{t'}, a^k_{t'}) + \bar{V}^{k-1}_{t+h}(s^k_{t+h}) \mid \mathcal{F}_{k-1}\right]. \tag{11}$$

Summing (11) for all $n \in \{0\} \cup [\frac{H}{h}]$, and using the linearity of expectation and the fact that $\bar{V}^k_{H+1}(s) = 0$ for all s, k, we have

$$\sum_{n=0}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^k_{nh+1}(s^k_{nh+1}) \mid \mathcal{F}_{k-1}\big] = \mathbb{E}\left[\sum_{t=1}^{H} r(s^k_t, a^k_t) \mid \mathcal{F}_{k-1}\right] + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^{k-1}_{nh+1}(s^k_{nh+1}) \mid \mathcal{F}_{k-1}\big]$$
$$\iff \bar{V}^k_1(s^k_1) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^k_{nh+1}(s^k_t) \mid \mathcal{F}_{k-1}\big] = \mathbb{E}\left[\sum_{t=1}^{H} r(s^k_t, a^k_t) \mid \mathcal{F}_{k-1}\right] + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^{k-1}_{nh+1}(s^k_{nh+1}) \mid \mathcal{F}_{k-1}\big]$$
$$\iff \bar{V}^k_1(s^k_1) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^k_{nh+1}(s^k_t) \mid \mathcal{F}_{k-1}\big] = V^{\pi_k}(s^k_1) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^{k-1}_{nh+1}(s^k_{nh+1}) \mid \mathcal{F}_{k-1}\big]$$
$$\iff \bar{V}^k_1(s^k_1) - V^{\pi_k}(s^k_1) = \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^{k-1}_{nh+1}(s^k_{nh+1}) - \bar{V}^k_{nh+1}(s^k_{nh+1}) \mid \mathcal{F}_{k-1}\big]. \tag{12}$$

The second line holds by the fact that s^k_1 is measurable w.r.t. $\mathcal{F}_{k-1}$ The third line holds since

$$V^{\pi_k}_1(s^k_1) = \mathbb{E}\left[\sum_{t=1}^{H} r(s^k_t, a^k_t) \mid s^k_1, \pi_k\right] = \mathbb{E}\left[\sum_{t=1}^{H} r(s^k_t, a^k_t) \mid \mathcal{F}_{k-1}\right].$$

Applying Lemma 15 from Appendix 14 with $g^k_t = \bar{V}^k_t$ for $t = nh + 1$, we obtain

$$(12) = \sum_{n=1}^{\frac{H}{h}-1} \sum_{s \in \mathcal{S}} \bar{V}^{k-1}_{nh+1}(s) - \mathbb{E}[\bar{V}^k_{nh+1}(s) \mid \mathcal{F}_{k-1}],$$

which concludes the proof. Note that the update of $\bar{V}^k_t$ occurs only at the visited state s^k_t and the update rule uses $\bar{V}^{k-1}_{t+h}$, i.e., it is measurable w.r.t. $\mathcal{F}_{k-1}$, and thus, it is valid to apply Lemma 15. □

Theorem 4 (Performance of h-RTDP). *Let $\epsilon, \delta > 0$. The following holds for h-RTDP:*

1. *With probability* $1 - \delta$*, for all* $K > 0$, $\quad \text{Regret}(K) \leq \frac{9SH(H-h)}{h} \ln(3/\delta)$.
2. $\Pr\left\{\exists \epsilon > 0 \ : \ N_\epsilon \geq \frac{9SH(H-h)\ln(3/\delta)}{h\epsilon}\right\} \leq \delta.$

Proof. We start by proving **Claim (1)**. We know that the following bounds hold on the regret:

$$\begin{aligned}\text{Regret}(K) &:= \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \overset{\text{(a)}}{\leq} \sum_{k=1}^{K} \bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \\ &\overset{\text{(b)}}{=} \sum_{k=1}^{K} \sum_{n=1}^{\frac{H}{h}-1} \sum_{s \in \mathcal{S}} \bar{V}_{nh+1}^{k-1}(s) - \mathbb{E}[\bar{V}_{nh+1}^k(s) \mid \mathcal{F}_{k-1}]. \end{aligned} \tag{13}$$

(a) is by the optimism of the value function (Lemma 2), and **(b)** is by Lemma 3.

We would like to show that (13) is the regret of a Decreasing Bounded Process (DBP). We start by defining

$$X_k := \sum_{n=1}^{\frac{H}{h}-1} \sum_{s \in \mathcal{S}} \bar{V}_{nh+1}^k(s). \tag{14}$$

We now prove that $\{X_k\}_{k \geq 0}$ is a DBP. Note that $\{X_k\}_{k \geq 0}$

1. is decreasing, since $\forall s, t,\ \bar{V}_t^k(s) \leq \bar{V}_t^{k-1}(s)$ by Lemma 2, and thus, their sum is also decreasing, and
2. is bounded since $\forall s, t\ \bar{V}_t^k(s) \geq V_t^*(s) \geq 0$ by Lemma 2, and thus, the sum is bounded from below by 0.

We can show that the initial value X_0 is also bounded as

$$X_0 = \sum_{n=1}^{\frac{H}{h}-1} \sum_{s \in \mathcal{S}} \bar{V}_{nh+1}^0(s) \leq \sum_{n=1}^{\frac{H}{h}-1} \sum_{s \in \mathcal{S}} H = \frac{SH(H-h)}{h}.$$

Using the linearity of expectation and the definition (14), we observe that (13) can be written as

$$\text{Regret}(K) \leq (13) = \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}],$$

which is regret of a DBP. Applying the bound on the regret of a DBP, Theorem 1, we conclude the proof of the first claim.

We now prove **Claim (2)**. Here we use a different technique than the one used in [15]. The technique allows us to prove uniform-PAC bounds for the approximate versions of h-RTDP described in Section 6. For these approximate versions, the uniform-PAC result is not a corollary of the regret bound and more careful analysis should be used.

For all $\epsilon > 0$, the following relations hold:

$$\begin{aligned}\mathbb{1}\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \epsilon\}\epsilon &\overset{\text{(a)}}{\leq} \mathbb{1}\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \epsilon\}\epsilon \\ &\overset{\text{(b)}}{\leq} \mathbb{1}\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \epsilon\}\left(\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k)\right) \\ &\overset{\text{(c)}}{=} \mathbb{1}\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \epsilon\}\left(\sum_{n=1}^{\frac{H}{h}-1} \sum_{s \in \mathcal{S}} \bar{V}_{nh+1}^{k-1}(s) - \mathbb{E}[\bar{V}_{nh+1}^k(s) \mid \mathcal{F}_{k-1}]\right) \\ &\overset{\text{(d)}}{=} \mathbb{1}\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \epsilon\}(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]). \end{aligned} \tag{15}$$

(a) holds since for all t, s, $\bar{V}_t^k(s) \geq V_t^*(s)$ by Lemma 2. **(b)** holds by the indicator function. **(c)** holds by Lemma 3. **(d)** holds by the definition of X_k from (14) and the linearity of expectation.

Let define $N_\epsilon(K) = \sum_{k=1}^{K} \mathbb{1}\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \ge \epsilon\}$ as the number of times $V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \ge \epsilon$ at the first K episodes. For all $\epsilon > 0$, we may write

$$N_\epsilon(K)\epsilon \overset{\text{(a)}}{=} \sum_{k=1}^{K} \mathbb{1}\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \ge \epsilon\}\epsilon \overset{\text{(b)}}{\le} \sum_{k=1}^{K} \mathbb{1}\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \ge \epsilon\}(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}])$$
$$\overset{\text{(c)}}{\le} \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}],$$

(a) holds by the definition of $N_\epsilon(K)$. **(b)** follows from (15). **(c)** holds because $\{X_k\}_{k\ge 0}$ is a DBP, and thus, $X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \ge 0$ a.s. Therefore, the following relation holds:

$$\left\{\forall K > 0 : \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \le \frac{9SH(H-h)}{h} \ln\frac{3}{\delta}\right\} \subseteq \left\{\forall \epsilon > 0 : N_\epsilon(K)\epsilon \le \frac{9SH(H-h)}{h} \ln\frac{3}{\delta}\right\},$$

from which we obtain that for any $K > 0$,

$$\Pr\left(\forall \epsilon > 0 : N_\epsilon(K)\epsilon \le \frac{9SH(H-h)}{h} \ln\frac{3}{\delta}\right) \ge \Pr\left(\forall K > 0 : \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \le \frac{9SH(Hh)}{h} \ln\frac{3}{\delta}\right) \overset{\text{(a)}}{\ge} 1 - \delta.$$

(a) holds because of the bound on the regret of DBP (see Theorem 1). Equivalently, for any $K > 0$,

$$\Pr\left(\exists \epsilon > 0 : N_\epsilon(K)\epsilon \ge \frac{9SH(H-h)}{h} \ln\frac{3}{\delta}\right) \le \delta. \tag{16}$$

Note that for all $\epsilon > 0$, $K_1 \ge K_2$, $\mathbb{1}\{N_\epsilon(K_2)\epsilon \ge C\} = 1$ implies $\mathbb{1}\{N_\epsilon(K_1)\epsilon \ge C\} = 1$, and thus, $\mathbb{1}\{N_\epsilon(K)\epsilon \ge C\} \le \lim_{K\to\infty} \mathbb{1}\{N_\epsilon(K)\epsilon \ge C\}$. Furthermore, $\mathbb{1}\{N_\epsilon(K)\epsilon \ge C\} \ge 0$ by definition. Thus, we can apply the Monotone Convergence Theorem to conclude the proof:

$$\Pr\left(\exists \epsilon > 0 : N_\epsilon\epsilon \ge \frac{9SH(H-h)}{h} \ln\frac{3}{\delta}\right) = \Pr\left(\lim_{K\to\infty}\left\{\exists \epsilon > 0 : N_\epsilon(K)\epsilon \ge \frac{9SH(H-h)}{h} \ln\frac{3}{\delta}\right\}\right)$$
$$= \mathbb{E}\left[\lim_{K\to\infty} \mathbb{1}\left\{\exists \epsilon > 0 : N_\epsilon(K)\epsilon \ge \frac{9SH(H-h)}{h} \ln\frac{3}{\delta}\right\}\right] \overset{\text{(a)}}{=} \lim_{K\to\infty} \mathbb{E}\left[\mathbb{1}\left\{\exists \epsilon > 0 : N_\epsilon(K)\epsilon \ge \frac{9SH(H-h)}{h} \ln\frac{3}{\delta}\right\}\right]$$
$$= \lim_{K\to\infty} \Pr\left(\exists \epsilon > 0 : N_\epsilon(K)\epsilon \ge \frac{9SH(H-h)}{h} \ln\frac{3}{\delta}\right) \overset{\text{(b)}}{\le} \delta.$$

(a) is by the Monotone Convergence Theorem by which $\mathbb{E}[\lim_{k\to\infty} X_k] = \lim_{k\to\infty} \mathbb{E}[X_k]$, for $X_k \ge 0$ and $X_k \le \lim_{k\to\infty} X_k$. **(b)** holds by (16). □

11 h-RTDP with Approximate Model

Algorithm 6 h-RTDP with Approximate Model (h-RTDP-AM)

init: $\forall s \in \mathcal{S},\ \forall n \in \{0\} \cup [\frac{H}{h}],\ \bar{V}^0_{nh+1}(s) = H - nh$
for $k \in [K]$ **do**
  Initialize s^k_1
  for $t \in [H]$ **do**
    if $(t-1) \mod h == 0$ **then**
      $h_c = t + h$
      $\bar{V}^k_t(s^k_t) = \hat{T}^h \bar{V}^{k-1}_{h_c}(s^k_t)$
    end if
    $a^k_t \in \arg\max_a r(s^k_t, a) + \hat{p}(\cdot|s^k_t, a)\hat{T}^{h_c - t - 1}\bar{V}^{k-1}_{h_c}$
    Act with a^k_t and observe $s^k_{t+1} \sim p(\cdot \mid s^k_t, a^k_t)$
  end for
end for

Algorithm 6 contains the pseudocode of h-RTDP with approximate model. The algorithm is exactly the same as h-RTDP (Algorithm 2) with the model p and optimal Bellman operator T replaced by their approximations $\hat{p}$ and $\hat{T}$. Meaning, h-RTDP is agnostic whether it uses the true or approximate model.

We now provide the full proofs of all results in Section 6.1 in their chronological order. We use the notation $\mathbb{E}_{\hat{P}}$ to denote expectation w.r.t. the approximate model, i.e., w.r.t. the dynamics $\hat{p}(s' \mid s, a)$ instead according to $p(s' \mid s, a)$.

Lemma 9. *For all $s \in \mathcal{S}$, $n \in \{0\} \cup [\frac{H}{h}]$, and $k \in [K]$:*

(i) Bounded / Optimism: $\hat{V}^*_{nh+1}(s) \le \bar{V}^k_{nh+1}(s)$.

(ii) Non-Increasing: $\bar{V}^k_{nh+1}(s) \le \bar{V}^{k-1}_{nh+1}(s)$.

Proof. Both claims are proven using induction.

(i) Let $n \in [0, \frac{H}{h} - 1]$ and denote $\hat{T}, \hat{V}^*$ as the optimal Bellman operators and optimal value of the approximate MDP $(\mathcal{S}, \mathcal{A}, \hat{p}, r, H)$. See that they satisfy usual Bellman equation 2.

By the initialization, $\forall s, t,\ \hat{V}^*_{1+hn}(s) \le V^0_{1+hn}(s)$. Assume the claim holds for $k-1$ episodes. Let s^k_t be the state the algorithm is at in the $t = 1 + hn$ time step of the k'th episode, i.e., at a time step in which a value update is taking place. By the value update of Algorithm 6,

$$\bar{V}^k_t(s^k_t) = (\hat{T}^h \bar{V}_{t+h})(s^k_t) \ge (\hat{T}^h \hat{V}^*_{t+h})(s^k_t) = \hat{V}^*_t(s^k_t).$$

The second relation holds by the induction hypothesis and the monotonicity of $\hat{T}^h$, a consequnce of the monotonicity of $\hat{T}$, the optimal Bellman operator [4]. The third relation holds by the recursion satisfied by the optimal value function (2). Thus, the induction step is proven for the first claim.

(ii) Let $n \in [0, \frac{H}{h} - 1]$ and let $t = 1 + hn$ be a time step in which a value update is taking place. To prove the base case of the second claim we use the optimistic initialization. Let s^1_t be the state the algorithm is at in the t'th time step of the first episode. By the update rule,

$$\begin{aligned}
\bar{V}^1_t(s^1_t) =\ & (\hat{T}^h \bar{V}^0_{t+h})(s^0_t) \\
\overset{(1)}{=} & \max_{\pi_0, \pi_1, .., \pi_{h-1}} \mathbb{E}_{\hat{P}'}[\sum_{t'=0}^{h-1} r(s'_t, \pi_{t'}(s'_t)) + \bar{V}^0_{t+h}(s_h) \mid s_0 = s^0_t] \\
\overset{(2)}{\le} & h + H - (t + h - 1) = H - (t-1) \overset{(3)}{=} \bar{V}^0_t(s^1_t).
\end{aligned}$$

Relation (1) is by the update rule (see Algorithm 6), when the expectation is taken place w.r.t. the approximate model $\hat{P}$. Relation (2) holds since $r(s,a) \in [0,1]$ and and by the optimistic initialization (see that for t the values at times step $t+h$ were not updated and keep their initial value). For (3) observe that $H-(t-1)$ is the value of the optimistic initialization.

Assume the second claim holds for $k-1$ episodes. Let s_t^k be the state that the algorithm is at in the t'th time step of the k'th episode. Again, assume that $t=1+hn$, a time step in which a value update is being done. By the value update of Algorithm 6, we have

$$\bar{V}_t^k(s_t^k) = (\hat{T}^h \bar{V}_{t+h}^{k-1})(s_t^k).$$

If s_t^k was previously updated, let $\bar{k}$ be the previous episode in which the update occured. By the induction hypothesis, we have that $\forall s,t,\ \bar{V}_t^{\bar{k}}(s) \geq \bar{V}_t^{k-1}(s)$. Using the monotonicity of T^h (due to the monotonicity of the Bellman operator),

$$(\hat{T}^h \bar{V}_{t+h}^{k-1})(s_t^k) \leq (\hat{T}^h \bar{V}_{t+h}^{\bar{k}})(s_t^k) = \bar{V}_t^{k-1}(s_t^k).$$

Thus, $\bar{V}_t^k(s_t^k) \leq \bar{V}^{k-1}(s_t^k)$ and the induction step is proved. If s_t^k was not previously updated, then $\bar{V}_t^{k-1}(s_t^k) = \bar{V}_t^0(s_t^k)$. In this case, the induction hypothesis implies that $\forall s', \bar{V}_{t+h}^{k-1}(s') \leq \bar{V}_{t+h}^0(s')$ and the result is proven similarly to the base case. □

Lemma 10. *The expected cumulative value update at the k'th episode of h-RTDP-AM satisfies the following relation:*

$$\begin{aligned}\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) = &\frac{H(H-1)}{2}\epsilon_P \\ &+ \sum_{n=1}^{\frac{H}{h}-1} \sum_{s\in\mathcal{S}} \bar{V}_{nh+1}^{k-1}(s) - \mathbb{E}[\bar{V}_{nh+1}^k(s) \mid \mathcal{F}_{k-1}].\end{aligned}$$

Proof. Let $n \in [0, \frac{H}{h}-1]$ and let $t = 1+hn$ be a time step in which a value update is taking place. By the definition of the update rule, the following holds for the update at the visited state s_t^k:

$$\begin{aligned}
&\bar{V}_t^k(s_t^k) = (\hat{T}^h \bar{V}_{t+h}^{k-1})(s_t^k) \qquad (17)\\
&= (\hat{T}^{\pi_k(t)} \cdots \hat{T}^{\pi_k(t+h-1)} \bar{V}_{t+h}^{k-1})(s_t^k)\\
&= \mathbb{E}_{P'}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \pi_k, s_t^k\right]\\
&= \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \pi_k, s_t^k\right]\\
&\quad + \mathbb{E}_{P'}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \pi_k, s_t^k\right] - \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \pi_k, s_t^k\right]\\
&= \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \pi_k, s_t^k\right]\\
&\quad + \sum_{t'=t}^{t+h-1} \sum_{s_{t'}} \Big(P^{\pi_k}(s_{t'} \mid s_t^k) - \hat{P}^{\pi_k}(s_{t'} \mid s_t^k)\Big) r(s_{t'}^k, a_{t'}^k) + \sum_{s_{t+h}} \Big(P^{\pi_k}(s_{t+h} \mid s_t^k) - \hat{P}^{\pi_k}(s_{t+h} \mid s_t^k)\Big) \bar{V}_{t+h}^{k-1}(s_{t+h}^k))\\
&\leq \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \pi_k, s_t^k\right]\\
&\quad + \sum_{t'=t}^{t+h-1} \sum_{s_{t'}} \Big|P^{\pi_k}(s_{t'} \mid s_t^k) - \hat{P}^{\pi_k}(s_{t'} \mid s_t^k)\Big| + (H-(t+h-1)) \sum_{s_{t+h}} \Big|P^{\pi_k}(s_{t+h} \mid s_t^k) - \hat{P}^{\pi_k}(s_{t+h} \mid s_t^k)\Big|.
\end{aligned}$$

Applying Lemma 16 we bound the above by,

$$
\begin{aligned}
(17) &\leq \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \pi_k, s_t^k\right] + \sum_{t'=t}^{t+h-1} (t'-t)\epsilon_P + (H-(t+h-1))h\epsilon_P \\
&= \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \pi_k, s_t^k\right] - \frac{1}{2}(h-1)h\epsilon_P + (H-t)h\epsilon_P \\
&= \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \mathcal{F}_{k-1}, s_t^k\right] - \frac{1}{2}(h-1)h\epsilon_P + (H-t)h\epsilon_P. \qquad (18)
\end{aligned}
$$

Where the second relation holds by using the close form of the arithmetic sum and by algebraic manipulations. For the third relation, we observe that given π_k, s_t^k the state $s_{t'}^k$ with $t' \geq t$ is independent of the past episodes (see 10),

$$
\mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \pi_k, s_t^k\right] = \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \mathcal{F}_{k-1}, s_t^k\right]
$$

Taking the conditional expectation w.r.t. $\mathcal{F}_{k-1}$ of both (17) and its RHS (18), using the tower property and the fact for all s, $\bar{V}_{H+1}(s) = 0$ we get,

$$
\begin{aligned}
\mathbb{E}\big[\bar{V}_t^k(s_t^k) \mid \mathcal{F}_{k-1}\big] \leq &\mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \mathcal{F}_{k-1}\right] \\
&- \frac{1}{2}(h-1)h\epsilon_P + (H-t)h\epsilon_P
\end{aligned}
$$

Let us denote $d_n := -\frac{1}{2}(h-1)h\epsilon_P + (H-n)h\epsilon_P$. Summing the above relation for all $n \in [\frac{H}{h}] - 1$, using linearity of expectation, and the fact $\bar{V}_{H+1}^k(s) =$ for all s, k,

$$
\sum_{n=0}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^k(s_t^k) \mid \mathcal{F}_{k-1}\big] = \mathbb{E}\left[\sum_{t=1}^{H} r(s_t^k, a_t^k) \mid \mathcal{F}_{k-1}\right] + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^{k-1}(s_{1+nh}^k) \mid \mathcal{F}_{k-1}\big] + \sum_{n=0}^{\frac{H}{h}-1} d_{1+nh} \qquad (19)
$$

By simple algebraic manipulation we get $\sum_{n=0}^{\frac{H}{h}-1} d_{1+nh} = \frac{1}{2}H(H-1)\epsilon_P$ (see Lemma 18). Thus, (19) has the following equivalent forms, by which we conclude the proof of this lemma.

$$
\begin{aligned}
&\iff \bar{V}_1^k(s_1^k) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^k(s_t^k) \mid \mathcal{F}_{k-1}\big] = \mathbb{E}\left[\sum_{t=1}^{H} r(s_t^k, a_t^k) \mid \mathcal{F}_{k-1}\right] + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^{k-1}(s_{1+nh}^k) \mid \mathcal{F}_{k-1}\big] + \frac{1}{2}H(H-1)\epsilon_P \\
&\iff \bar{V}_1^k(s_1^k) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^k(s_t^k) \mid \mathcal{F}_{k-1}\big] = V^{\pi_k}(s_1^k) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^{k-1}(s_{1+nh}^k) \mid \mathcal{F}_{k-1}\big] + \frac{1}{2}H(H-1)\epsilon_P \\
&\iff \bar{V}_1^k(s_1^k) - V^{\pi_k}(s_1^k) = \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^{k-1}(s_{1+nh}^k) - \bar{V}_{1+nh}^k(s_{1+nh}^k) \mid \mathcal{F}_{k-1}\big] + \frac{1}{2}H(H-1)\epsilon_P \\
&\iff \bar{V}_1^k(s_1^k) - V^{\pi_k}(s_1^k) = \sum_{k=1}^{K} \sum_{n=1}^{\frac{H}{h}-1} \sum_{s} \bar{V}_{nh+1}^{k-1}(s) - \mathbb{E}[\bar{V}_{nh+1}^k(s) \mid \mathcal{F}_{k-1}] + \frac{1}{2}H(H-1)\epsilon_P
\end{aligned}
$$

The second line holds by the fact s_1^k is measurable w.r.t. $\mathcal{F}_{k-1}$, the third line holds since

$$
V_1^{\pi_k}(s_1^k) = \mathbb{E}\left[\sum_{t=1}^{H} r(s_t^k, a_t^k) \mid \mathcal{F}_{k-1}\right].
$$

The forth line holds by Lemma 15 with $\bar{V}_t^k = g_t^k$ for $t = nh + 1$. See that the update of $\bar{V}_t^k$ occurs only at the visited state s_t^k and the update rule uses $\bar{V}_{t+1}^{k-1}$, i.e., it is measurable w.r.t. to $\mathcal{F}_{k-1}$, and it is valid to apply the lemma.

□

Theorem 5 (Performance of h-RTDP-AM)**.** *Let $\epsilon, \delta > 0$. The following holds for h-RTDP-AM:*

1. *With probability $1-\delta$, for all $K > 0$,* $\text{Regret}(K) \le \frac{9SH(H-h)}{h}\ln(3/\delta) + H(H-1)\epsilon_P K$.
2. *Let $\Delta_P = H(H-1)\epsilon_P$. Then,* $\Pr\left\{\exists \epsilon > 0 \ : \ N_\epsilon^{\Delta_P} \ge \frac{9SH(H-h)\ln(3/\delta)}{h\epsilon}\right\} \le \delta$.

Proof. We start by proving **claim (1)**. The following bounds on the regret hold.

$$\begin{aligned}
\text{Regret}(K) &:= \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \\
&\le \sum_{k=1}^{K} \hat{V}_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) + \frac{H(H-1)}{2}\epsilon_P \\
&\le \sum_{k=1}^{K} \bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) + \frac{H(H-1)}{2}\epsilon_P \\
&= H(H-1)\epsilon_P K + \sum_{k=1}^{K}\sum_{n=1}^{\frac{H}{h}-1}\sum_{s} \bar{V}_{nh+1}^{k-1}(s) - \mathbb{E}[\bar{V}_{nh+1}^{k}(s) \mid \mathcal{F}_{k-1}]
\end{aligned} \tag{20}$$

The second relation holds by Lemma 17 which relates the optimal value of the approximate model to the optimal value of the environment. The third relation is by the optimism of the value function (Lemma 9), and the forth relation is by Lemma 10.

We now observe the regret is a regret of a Decreasing Bounded Process. Let

$$X_k := \sum_{n=1}^{\frac{H}{h}-1}\sum_{s} \bar{V}_{nh+1}^{k}(s), \tag{21}$$

and observe that $\{X_k\}_{g\ge 0}$ is a Decreasing Bounded Process.

1. It is decreasing since for all s, t $\bar{V}_t^k(s) \le \bar{V}_t^{k-1}(s)$ by Lemma 9. Thus, their sum is also decreasing.
2. It is bounded since for all s, t $\bar{V}_t^k(s) \ge V_t^*(s) \ge 0$ by Lemma 9. Thus, the sum is bounded from below by 0.

See that the initial value can be bounded as follows,

$$X_0 = \sum_{n=1}^{\frac{H}{h}-1}\sum_{s} \bar{V}_{nh+1}^{0}(s) \le \sum_{n=1}^{\frac{H}{h}-1}\sum_{s} H = \frac{SH(H-h)}{h}.$$

Using linearity of expectation and the definition (21) we observe that (20) can be written,

$$\text{Regret}(K) \le (20) = H(H-1)\epsilon_P K + \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}],$$

which is regret of A Bounded Decreasing Process. Applying the regret bound on DBP, Theorem 1, we conclude the proof of the first claim.

We now prove the **claim (2)** using the proving technique at Theorem 4. Denote $\Delta_P = H(H-1)\epsilon_P$. The following relations hold for all $\epsilon > 0$.

$$
\begin{aligned}
&\mathbb{1}\left\{\hat{V}_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_P}{2} + \epsilon\right\}\left(\epsilon + \frac{\Delta_P}{2}\right) \\
&\leq \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_P}{2} + \epsilon\right\}\left(\epsilon + \frac{\Delta_P}{2}\right) \\
&\leq \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_P}{2} + \epsilon\right\}\left(\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k)\right) \\
&= \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_P}{2} + \epsilon\right\}\left(\sum_{n=1}^{\frac{H}{h}-1}\sum_s \bar{V}_{nh+1}^{k-1}(s) - \mathbb{E}[\bar{V}_{nh+1}^k(s) \mid \mathcal{F}_{k-1}] + \frac{\Delta_P}{2}\right) \\
&= \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_P}{2} + \epsilon\right\}\left(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] + \frac{\Delta_P}{2}\right).
\end{aligned}
$$

The first relation holds since for all t, s, $\bar{V}_t^k(s) \geq \hat{V}_t^*(s)$ by Lemma 9. The second relation holds by the indicator function and the third relation holds by Lemma 10. The forth relation holds by the definition of X_k (21) and linearity of expectation. Using an algebraic manipulation the above leads to the following relation,

$$
\mathbb{1}\left\{\hat{V}_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_P}{2} + \epsilon\right\}\epsilon \leq \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_P}{2} + \epsilon\right\}(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]) \tag{22}
$$

As we wish the final performance to be compared to V^* and not $\hat{V}$ we use the the first claim of Lemma 17, by which for all s, $\hat{V}_1^*(s) \geq V_1^*(s) - \frac{\Delta_P}{2}$. This implies that

$$
\mathbb{1}\left\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \Delta_P + \epsilon\right\} \leq \mathbb{1}\left\{\hat{V}_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_P}{2} + \epsilon\right\}. \tag{23}
$$

Combining all the above, we get

$$
\begin{aligned}
&\mathbb{1}\left\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \Delta_P + \epsilon\right\}\epsilon \\
&\leq \mathbb{1}\left\{\hat{V}_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_P}{2} + \epsilon\right\}\epsilon \\
&\leq \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_P}{2} + \epsilon\right\}(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]).
\end{aligned} \tag{24}
$$

The first relation is by (23) and the second relation by (22).

Define $N_\epsilon(K) = \sum_{k=1}^K \mathbb{1}\left\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \Delta_P + \epsilon\right\}$ as the number of times $V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \Delta_P + \epsilon$ at the first K episodes. Summing the above inequality (24) for all $k \in [K]$ and denote we get that for all $\epsilon > 0$

$$
\begin{aligned}
N_\epsilon(K)\epsilon &= \sum_{k=1}^K \mathbb{1}\left\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \Delta_P + \epsilon\right\}\epsilon \\
&\leq \sum_{k=1}^K \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_P}{2} + \epsilon\right\}(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]) \\
&\leq \sum_{k=1}^K X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}].
\end{aligned}
$$

The first relation holds by definition, the second by (24) and the third relation holds as $\{X_k\}_{k\geq 0}$ is a DBP (21) and, thus, $X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \geq 0$ a.s. . Thus, the following relation holds

$$
\left\{\forall K > 0 : \sum_{k=1}^K X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \leq \frac{9SH(H-h)}{h} \ln\frac{3}{\delta}\right\} \subseteq \left\{\forall \epsilon > 0 : N_\epsilon(K)\epsilon \leq \frac{9SH(H-h)}{h} \ln\frac{3}{\delta}\right\},
$$

from which we get that for any $K > 0$

$$\Pr\left(\forall \epsilon > 0 : N_\epsilon(K)\epsilon \le \frac{9SH(H-h)}{h} \ln \frac{3}{\delta}\right)$$
$$\ge \Pr\left(\forall K > 0 : \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \le \frac{9SH(Hh)}{h} \ln \frac{3}{\delta}\right) \ge 1 - \delta,$$

and the third relation holds the bound on the regret of DBP, Theorem 1. Equivalently, for any $K > 0$,

$$\Pr\left(\exists \epsilon > 0 : N_\epsilon(K)\epsilon \ge \frac{9SH(H-h)}{h} \ln \frac{3}{\delta}\right) \le \delta. \tag{25}$$

Applying the Monotone Convergence Theorem as in the proof of Theorem 4 we conclude the proof.

□

Algorithm 7 h-RTDP with Approximate Value Updates (h-RTDP-AV)

init: $\forall s \in \mathcal{S},\ n \in \{0\} \cup [\frac{H}{h}],\ \bar{V}^0_{nh+1}(s) = H - nh$
for $k \in [K]$ **do**
  Initialize s^k_1
  for $t \in [H]$ **do**
    if $(t-1) \mod h == 0$ **then**
      $h_c = t + h$
      $\bar{V}^k_t(s^k_t) = \epsilon_V(s^k_t) + T^h \bar{V}^{k-1}_{h_c}(s^k_t)$; $\bar{V}^k_t(s^k_t) \leftarrow \min\{\bar{V}^k_t(s^k_t), \bar{V}^{k-1}_t(s^k_t)\}$;
    end if
    $a^k_t \in \arg\max_a r(s^k_t, a) + p(\cdot|s^k_t, a) T^{h_c - t - 1} \bar{V}^{k-1}_{h_c}$;
    Act with a^k_t and observe $s^k_{t+1} \sim p(\cdot \mid s^k_t, a^k_t)$
  end for
end for

12 h-RTDP with Approximate Value updates

Lemma 11. *For all $s \in \mathcal{S}$, $n \in \{0\} \cup [\frac{H}{h}]$, and $k \in [K]$:*

(i) Bounded / Optimism:

$$V^*_{nh+1}(s) \leq \bar{V}^k_{nh+1}(s) + \epsilon_V(\frac{H}{h} - n).$$

(ii) Non-Increasing: $\bar{V}^k_{nh+1}(s) \leq \bar{V}^{k-1}_{nh+1}(s)$.

Proof. We prove the first claim by induction. The second claim holds by construction.

(i) Let $n \in \{0\} \cup [\frac{H}{h}]$. By the optimistic initialization, $\forall s, n,\ V^*_{1+hn}(s) - \epsilon_V(\frac{H}{h} - n) \leq V^*_{1+hn}(s) \leq V^0_{1+hn}(s)$. Assume the claim holds for $k-1$ episodes. Let s^k_t be the state the algorithm is at in the $t = 1 + hn$ time step of the k'th episode, i.e., at a time step in which a value update is taking place. Let $e \in \mathbb{R}^S$ be the constant vector of ones. By the value update of Algorithm 7,

$$\bar{V}^k_t(s^k_t) = \min\{\epsilon_V(s^k_t) + T^h \bar{V}^{k-1}_{h_c}(s^k_t), \bar{V}^{k-1}_t(s^k_t)\}. \tag{26}$$

If the minimal value is $\bar{V}^{k-1}_t(s^k_t)$ then $\bar{V}^k_t(s^k_t)$ satisfies the induction hypothesis by the induction assumption. If $\epsilon_V(s^k_t) + T^h \bar{V}^{k-1}_{h_c}(s^k_t)$ is the minimal value in (26), then the following relation holds,

$$\begin{aligned}
\bar{V}^k_t(s^k_t) &= \epsilon_V(s^k_t) + T^h \bar{V}^{k-1}_{t+h}(s^k_t) \\
&\geq -\epsilon_V + T^h \bar{V}^{k-1}_{t+h}(s^k_t) \\
&\geq -\epsilon_V + T^h \left(V^*_{t+h} - e\epsilon_V(\frac{H}{h} - n - 1)\right)(s^k_t) \\
&= -\epsilon_V + T^h V^*_{t+h}(s^k_t) - \epsilon_V(\frac{H}{h} - n - 1) \\
&= T^h V^*_{t+h}(s^k_t) - \epsilon_V(\frac{H}{h} - n) \\
&= V^*_t(s^k_t) - \epsilon_V(\frac{H}{h} - n).
\end{aligned}$$

The second relation holds by the assumption $|\epsilon_V(s^k_t)| \leq \epsilon_V$. The third relation by the induction hypothesis and the monotonicity of T^h. The forth relation holds since for any constant $\alpha \in \mathbb{R}$ and $V \in \mathbb{R}^s$, $T(V + \alpha e) = TV + \alpha$ (e.g.,[4]) and thus $T^h(V + \alpha e) = T^h V + \alpha$. Lastly, the fifth relation holds by the Bellman equations (2).

(ii) The second claim holds by construction of the update rule $\bar{V}^k_t(s^k_t) \leftarrow \min\{\bar{V}^k_t(s^k_t), \bar{V}^{k-1}_t(s^k_t)\}$ which enforces $\bar{V}^k_t(s) \leq \bar{V}^{k-1}_t(s)$ for every updated state, and thus for all s and t. □

Lemma 12. *The expected cumulative value update at the k'th episode of h-RTDP-AV satisfies the following relation:*

$$\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k)$$
$$\leq \frac{H}{h}\epsilon_V + \sum_{k=1}^{K}\sum_{n=1}^{\frac{H}{h}-1}\sum_{s\in\mathcal{S}} \bar{V}_{nh+1}^{k-1}(s) - \mathbb{E}[\bar{V}_{nh+1}^{k}(s) \mid \mathcal{F}_{k-1}].$$

Proof. Let $n \in \{0\} \cup [\frac{H}{h} - 1]$ and let $t = 1 + hn$ be a time step in which a value update is taking place. By the definition of the update rule, the following holds for the update at the visited state s_t^k:

$$\begin{aligned}
\bar{V}_t^k(s_t^k) &= \epsilon_V(s_t^k) + (T^h \bar{V}_{t+h}^{k-1})(s_t^k) \\
&\leq \epsilon_V + (T^{\pi_k(t)} \cdots T^{\pi_k(t+h-1)} \bar{V}_{t+h}^{k-1})(s_t^k) \\
&= \epsilon_V + \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \mathcal{F}_{k-1}, s_t^k\right].
\end{aligned}$$

Where the third relation holds by the same argument as in (10). Taking the conditional expectation w.r.t. $\mathcal{F}_{k-1}$, using the tower property and the fact for all s, $\bar{V}_{H+1}(s) = 0$ we get,

$$\mathbb{E}\big[\bar{V}_t^k(s_t^k) \mid \mathcal{F}_{k-1}\big] \leq \epsilon_V + \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(s_{t+h}^k) \mid \mathcal{F}_{k-1}\right].$$

Summing the above relation for all $n \in \{0\} \cup [\frac{H}{h} - 1]$, using linearity of expectation, and the fact $\bar{V}_{H+1}^k(s) =$ for all s, k,

$$\begin{aligned}
&\sum_{n=0}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^k(s_t^k) \mid \mathcal{F}_{k-1}\big] \leq \frac{H}{h}\epsilon_V + \mathbb{E}\left[\sum_{t=1}^{H} r(s_t^k, a_t^k) \mid \mathcal{F}_{k-1}\right] + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^{k-1}(s_{1+nh}^k) \mid \mathcal{F}_{k-1}\big] \\
\iff &\bar{V}_1^k(s_1^k) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^k(s_t^k) \mid \mathcal{F}_{k-1}\big] \leq \frac{H}{h}\epsilon_V + \mathbb{E}\left[\sum_{t=1}^{H} r(s_t^k, a_t^k) \mid \mathcal{F}_{k-1}\right] + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^{k-1}(s_{1+nh}^k) \mid \mathcal{F}_{k-1}\big] \\
\iff &\bar{V}_1^k(s_1^k) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^k(s_t^k) \mid \mathcal{F}_{k-1}\big] \leq \frac{H}{h}\epsilon_V + V^{\pi_k}(s_1^k) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^{k-1}(s_{1+nh}^k) \mid \mathcal{F}_{k-1}\big] \\
\iff &\bar{V}_1^k(s_1^k) - V^{\pi_k}(s_1^k) \leq \frac{H}{h}\epsilon_V + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}_{1+nh}^{k-1}(s_{1+nh}^k) - \bar{V}_{1+nh}^k(s_{1+nh}^k) \mid \mathcal{F}_{k-1}\big] \\
\iff &\bar{V}_1^k(s_1^k) - V^{\pi_k}(s_1^k) \leq \frac{H}{h}\epsilon_V + \sum_{k=1}^{K}\sum_{n=1}^{\frac{H}{h}-1}\sum_{s} \bar{V}_{nh+1}^{k-1}(s) - \mathbb{E}[\bar{V}_{nh+1}^k(s) \mid \mathcal{F}_{k-1}]
\end{aligned}$$

The second line holds by the fact s_1^k is measurable w.r.t. $\mathcal{F}_{k-1}$, and the third line holds since

$$V_1^{\pi_k}(s_1^k) = \mathbb{E}\left[\sum_{t=1}^{H} r(s_t^k, a_t^k) \mid \mathcal{F}_{k-1}\right].$$

The fifth line holds by by Lemma 15 with $\bar{V}_t^k = g_t^k$ for $t = nh + 1$. See that the update of $\bar{V}_t^k$ occurs only at the visited state s_t^k and the update rule uses $\bar{V}_{t+1}^{k-1}$, i.e., it is measurable w.r.t. to $\mathcal{F}_{k-1}$, and it is valid to apply the lemma.

□

Theorem 6 (Performance of h-RTDP-AV)**.** *Let $\epsilon, \delta > 0$. The following holds for h-RTDP-AV:*

1. *With probability $1 - \delta$, for all $K > 0$,* $\;\mathrm{Regret}(K) \leq \frac{9SH(H-h)}{h}(1 + \frac{H}{h}\epsilon_V)\ln(\frac{3}{\delta}) + \frac{2H}{h}\epsilon_V K.$
2. *Let $\Delta_V = 2H\epsilon_V$. Then,* $\;\Pr\left\{\exists \epsilon > 0 \;:\; N_\epsilon^{\frac{\Delta_V}{h}} \geq \frac{9SH(H-h)(1+\frac{\Delta_V}{2h})\ln(\frac{3}{\delta})}{h\epsilon}\right\} \leq \delta.$

Proof. We start by proving **claim (1)**. The following bounds on the regret hold.

$$\begin{aligned}
\mathrm{Regret}(K) &:= \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \\
&\leq \sum_{k=1}^{K} \bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) + \frac{H}{h}\epsilon_V \\
&= \frac{2H}{h}\epsilon_V K + \sum_{k=1}^{K}\sum_{n=1}^{\frac{H}{h}-1}\sum_s \bar{V}_{nh+1}^{k-1}(s) - \mathbb{E}[\bar{V}_{nh+1}^k(s) \mid \mathcal{F}_{k-1}] \qquad (27)
\end{aligned}$$

The second relation is by the approximated optimism of the value function when approximate value updates are used (Lemma 11). The third relation is by Lemma 12.

We now observe the regret is a regret of a Decreasing Bounded Process. Let

$$X_k := \sum_{n=1}^{\frac{H}{h}-1}\sum_s \bar{V}_{nh+1}^k(s), \qquad (28)$$

and observe that $\{X_k\}_{g\geq 0}$ is a Decreasing Bounded Process.

1. It is decreasing since for all s, t $\bar{V}_t^k(s) \leq \bar{V}_t^{k-1}(s)$ by Lemma 11. Thus, their sum is also decreasing.
2. It is bounded since for all $s, n \in [\frac{H}{h}] - 1$,

   $$\bar{V}_{1+hn}^k(s) \geq V_{1+hn}^*(s) - \epsilon_V(\frac{H}{h} - n) \geq -\epsilon_V(\frac{H}{h} - n) \geq -\epsilon_V\frac{H}{h}$$

   by Lemma 11. Thus, X_0 which is a sum of the above terms is bounded from below by $-\frac{\epsilon_V}{h}\frac{SH(H-h)}{h}$.

See that the initial value can be bounded as follows,

$$X_0 = \sum_{n=1}^{\frac{H}{h}-1}\sum_s \bar{V}_{nh+1}^0(s) \leq \sum_{n=1}^{\frac{H}{h}-1}\sum_s H = \frac{SH(H-h)}{h}.$$

Using linearity of expectation and the definition (14) we observe that (27) can be written,

$$\mathrm{Regret}(K) \leq (27) = \frac{2H}{h}\epsilon_V K + \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}],$$

which is regret of A Bounded Decreasing Process. Applying the regret bound on DBP, Theorem 1 we conclude the proof of the first claim.

We now prove **claim (2)** using the proving technique at Theorem 4. Denote $\Delta_V = 2H\epsilon_V$. The following relations hold for all $\epsilon > 0$.

$$\begin{aligned}
&\mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{2h} + \epsilon\right\}\left(\epsilon + \frac{\Delta_V}{2h}\right) \\
&\leq \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{2h} + \epsilon\right\}\left(\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k)\right) \\
&= \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{2h} + \epsilon\right\}\left(\sum_{n=1}^{\frac{H}{h}-1}\sum_s \bar{V}_{nh+1}^{k-1}(s) - \mathbb{E}[\bar{V}_{nh+1}^k(s) \mid \mathcal{F}_{k-1}] + \frac{\Delta_V}{2h}\right) \\
&= \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{2h} + \epsilon\right\}\left(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] + \frac{\Delta_V}{2h}\right).
\end{aligned}$$

The first relation holds by the indicator function and the second relation by Lemma 12. The third relation holds by the definition of X_k (28) and linearity of expectation. Using an algebraic manipulation the above leads to the following relation,

$$\mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{2h} + \epsilon\right\}\epsilon \leq \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{2h} + \epsilon\right\}(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]) \tag{29}$$

As we wish the final performance to be compared to V^* we use the the first claim of Lemma 11, by which for all s, k, $\bar{V}_1^k(s) \geq V_1^*(s) - \frac{\Delta_V}{2h}$. This implies that

$$\mathbb{1}\left\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{h} + \epsilon\right\} \leq \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{2h} + \epsilon\right\}. \tag{30}$$

Combining the above we get

$$\begin{aligned}
&\mathbb{1}\left\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{h} + \epsilon\right\}\epsilon \\
&\leq \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{2h} + \epsilon\right\}\epsilon \\
&\leq \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{2h} + \epsilon\right\}(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]).
\end{aligned} \tag{31}$$

The first relation is by (30) and the second relation by (29).

Define $N_\epsilon(K) = \sum_{k=1}^K \mathbb{1}\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{h} + \epsilon\}$ as the number of times $V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{h} + \epsilon$ at the first K episodes. Summing the above inequality (31) for all $k \in [K]$ and denote we get that for all $\epsilon > 0$

$$\begin{aligned}
N_\epsilon(K)\epsilon &= \sum_{k=1}^K \mathbb{1}\left\{V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{h} + \epsilon\right\}\epsilon \\
&\leq \sum_{k=1}^K \mathbb{1}\left\{\bar{V}_1^k(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_V}{2h} + \epsilon\right\}(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]) \\
&\leq \sum_{k=1}^K X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}].
\end{aligned}$$

The first relation holds by definition, the second by (31) and the third relation holds as $\{X_k\}_{k\geq 0}$ is a DBP (28) and, thus, $X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \geq 0$ a.s. . Thus, the following relation holds

$$\begin{aligned}
&\left\{\forall K > 0 : \sum_{k=1}^K X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \leq \frac{9SH(H-h)}{h}(1 + \frac{H}{h}\epsilon_V)\ln\frac{3}{\delta}\right\} \\
&\subseteq \left\{\forall \epsilon > 0 : N_\epsilon(K)\epsilon \leq \frac{9SH(H-h)}{h}(1 + \frac{H}{h}\epsilon_V)\ln\frac{3}{\delta}\right\},
\end{aligned}$$

from which we get for any $K > 0$

$$\begin{aligned}
&\Pr\left(\forall \epsilon > 0 : N_\epsilon(K)\epsilon \leq \frac{9SH(H-h)}{h}(1 + \frac{H}{h}\epsilon_V)\ln\frac{3}{\delta}\right) \\
&\geq \Pr\left(\forall K > 0 : \sum_{k=1}^K X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \leq \frac{9SH(Hh)}{h}(1 + \frac{H}{h}\epsilon_V)\ln\frac{3}{\delta}\right) \geq 1 - \delta,
\end{aligned}$$

and the third relation holds the bound on the regret of DBP, Theorem 1. Equivalently, for any $K > 0$,

$$\Pr\left(\exists \epsilon > 0 : N_\epsilon(K)\epsilon \geq \frac{9SH(H-h)}{h}(1 + \frac{H}{h}\epsilon_V)\ln\frac{3}{\delta}\right) \leq \delta. \tag{32}$$

Applying the Monotone Convergence Theorem as in the proof of Theorem 4 we conclude the proof.

□

Algorithm 8 h-RTDP with Approximate State Abstraction (h-RTDP-AA)

init: $\forall s_\phi \in \mathcal{S}_\phi,\ n \in \{0\} \cup [\frac{H}{h}],\ \bar{V}^0_{\phi,nh+1}(s_\phi) = H - nh$
for $k \in [K]$ **do**
  Initialize s^k_1
  for $t \in [H]$ **do**
    if $(t-1) \mod h == 0$ **then**
      $h_c = t + h$; $\bar{V}^k_{\phi,t}(\phi_t(s^k_t)) = T^h_\phi \bar{V}^{k-1}_{\phi,h_c}(s^k_t)$;
      $\bar{V}^k_{\phi,t}(\phi_t(s^k_t)) \leftarrow \min\Big\{\bar{V}^k_{\phi,t}(\phi_t(s^k_t)), \bar{V}^{k-1}_{\phi,t}(\phi_t(s^k_t))\Big\}$;
    end if
    $a^k_t \in \arg\max_a r(s^k_t, a) + p(\cdot|s^k_t, a) T^{h_c-t-1}_\phi \bar{V}^{k-1}_{\phi,h_c}$;
    Act with a^k_t and observe $s^k_{t+1} \sim p(\cdot \mid s^k_t, a^k_t)$
  end for
end for

13 h-RTDP with Approximate State Abstraction

In this section we analyze the performance of h-RTDP performance which uses approximate abstraction. For clarity we restate the assumption we make on the approximate abstraction and the definition of equivalent set under abstraction.

Assumption 1 (Approximate Abstraction, [23], definition 3.3)**.** *For any $s, s' \in \mathcal{S}$ and $n \in \{0\} \cup [\frac{H}{h} - 1]$ for which $\phi_{nh+1}(s) = \phi_{nh+1}(s')$, we have $|V^*_{nh+1}(s) - V^*_{nh+1}(s')| \leq \epsilon_A$.*

An important quantity in our analysis is the set of states equivalent to a given state s under ϕ_{nh+1}.

Definition 1 (Equivalent Set Under Abstraction)**.** *For any $s \in \mathcal{S}$ and $n \in \{0\} \cup [\frac{H}{h} - 1]$, we define the set of states equivalent to s under ϕ_{nh+1} as $\Phi_{nh+1}(s) := \{s' \in \mathcal{S} : \phi_{nh+1}(s) = \phi_{nh+1}(s')\}$.*

Before we supply with the proof we emphasize an important difference in the definition of the value function $\bar{V}^k_t$ when using abstraction. Unlike the usual definition of $\bar{V}^k_t : \mathcal{S} \to \mathbb{R}$, in case of abstraction $\bar{V}^k_{\phi,t}$ is a mapping from the *abstract state space* to the reals, i.e., $\bar{V}^k_{\phi,t} : \mathcal{S}_\phi \to \mathbb{R}$. Meaning, $\bar{V}^k_{\phi,t}$ is defined on the abstract state space. Given a state $s \in \mathcal{S}$ we need to query ϕ_t to obtain its value at time t by $\bar{V}^k_t(\phi_t(s))$.

Lemma 13. *For all $s \in \mathcal{S}$, $n \in \{0\} \cup [\frac{H}{h}]$, and $k \in [K]$:*

(i) Optimism:

$$\max_{s' \in \Phi_{nh+1}(s)} V^*_{nh+1}(s') \leq \bar{V}^k_{nh+1}(\phi_{nh+1}(s)) + \epsilon_A(\frac{H}{h} - n).$$

(ii) Bounded: $\bar{V}^k_{nh+1}(\phi_{nh+1}(s)) \geq 0$.

(iii) Non-Increasing: $\bar{V}^k_{nh+1}(\phi_{nh+1}(s)) \leq \bar{V}^{k-1}_{nh+1}(\phi_{nh+1}(s))$.

Proof. We prove the first claim by induction. The second and third claims hold by construction.

(i) Let $n \in \{0\} \cup [\frac{H}{h} - 1]$. By the optimistic initialization, $\forall s, n,\ V^*_{1+hn}(s) - \epsilon_A(\frac{H}{h} - n) \leq V^*_{1+hn}(s) \leq V^0_{1+hn}(\phi_{1+hn}(s))$. Assume the claim holds for $k-1$ episodes. Let s^k_t be the state the algorithm is at in the $t = 1 + hn$ time step of the k'th episode, i.e., at a time step in which a value update is taking place. By the value update of Algorithm 8,

$$\bar{V}^k_t(\phi(s^k_t)) = \min\big\{T^h \bar{V}^{k-1}_{h_c}(s^k_t), \bar{V}^{k-1}_t(\phi(s^k_t))\big\}. \tag{33}$$

If the minimal value is $\bar{V}^{k-1}_t(\phi(s^k_t))$ then $\bar{V}^k_t(\phi(s^k_t))$ satisfies the induction hypothesis by the induction assumption. If $T^h \bar{V}^{k-1}_{h_c}(s^k_t)$ is the minimal value in (33), then the following relation holds,

$$
\begin{aligned}
\bar{V}_t^k(\phi_t(s_t^k)) &= \max_{\pi_0,\pi_1,..,\pi_{h-1}} \mathbb{E}[\sum_{t'=0}^{h-1} r(s_t', \pi_{t'}(s_t')) + \bar{V}_{t+h}^{k-1}(\phi(s_h)) \mid s_0 = s_t^k] \\
&\geq \max_{\pi_0,\pi_1,..,\pi_{h-1}} \mathbb{E}[\sum_{t'=0}^{h-1} r(s_t', \pi_{t'}(s_t')) + \max_{s' \in \Phi_{t+h}(s_h)} V_{t+h}^*(s') - \epsilon_A\left(\frac{H}{h} - n - 1\right) \mid s_0 = s_t^k] \\
&= \max_{\pi_0,\pi_1,..,\pi_{h-1}} \mathbb{E}[\sum_{t'=0}^{h-1} r(s_t', \pi_{t'}(s_t')) + \max_{s' \in \Phi_{t+h}(s_h)} V_{t+h}^*(s') \mid s_0 = s_t^k] - \epsilon_A\left(\frac{H}{h} - n - 1\right) \\
&\geq \max_{\pi_0,\pi_1,..,\pi_{h-1}} \mathbb{E}[\sum_{t'=0}^{h-1} r(s_t', \pi_{t'}(s_t')) + V_{t+h}^*(s_h) \mid s_0 = s_t^k] - \epsilon_A\left(\frac{H}{h} - n - 1\right) \\
&= V_t^*(s_t^k) - \epsilon_A\left(\frac{H}{h} - n - 1\right) \\
&\geq \max_{s' \in \Phi_t(s_t^k)} V_t^*(s_t^k) - \epsilon_A - \epsilon_A\left(\frac{H}{h} - n - 1\right) \\
&= \max_{s' \in \Phi_t(s_t^k)} V_t^*(s_t^k) - \epsilon_A\left(\frac{H}{h} - n\right).
\end{aligned}
$$

The first relation is the definition of the update rule. The second relation holds by the monotonicity of the $\max$ operator together with the induction assumption. The third relation as the extracted term out of the $\max$ is constant. The forth relation holds by the definition of the $\max$ operation. The fifth relation by the Bellman equations V_t^* satisfies (2), and the sixth relation by Assumption 1.

(ii) The second claim holds by construction of the update rule $\bar{V}_t^k(s_t^k) \leftarrow \min\{\bar{V}_t^k(s_t^k), \bar{V}_t^{k-1}(s_t^k)\}$ which enforces $\bar{V}_t^k(s) \leq \bar{V}_t^{k-1}(s)$ for every updated state, and thus for all s and t.

(iii) The third claim holds since $V_t^k(\phi_t(s))$ is initialized with positive elements and is updated by itself and positive elements, as $r(s,a) \geq 0$. Thus, it remains positive a.s. . □

Lemma 14. *The expected cumulative value update at the k'th episode of h-RTDP-AA satisfies the following relation:*

$$
\begin{aligned}
&\bar{V}_1^k(\phi(s_1^k)) - V_1^{\pi_k}(s_1^k) \\
&\leq \sum_{k=1}^{K} \sum_{n=1}^{\frac{H}{h}-1} \sum_{s_\phi \in \mathcal{S}_\phi} \bar{V}_{nh+1}^{k-1}(s_\phi) - \mathbb{E}[\bar{V}_{nh+1}^k(s_\phi) \mid \mathcal{F}_{k-1}].
\end{aligned}
$$

Proof. Let $n \in \{0\} \cup [\frac{H}{h} - 1]$ and let $t = 1 + hn$ be a time step in which a value update is taking place. By the definition of the update rule, the following holds for the update at the visited state s_t^k:

$$
\begin{aligned}
\bar{V}_t^k(\phi_t(s_t^k)) &\leq \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(\phi_{t+h}(s_{t+h}^k)) \mid \pi_k, s_t^k\right] \\
&= \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(\phi_{t+h}(s_{t+h}^k)) \mid \mathcal{F}_{k-1}, s_t^k\right]
\end{aligned}
$$

where the last relation follows by the same argument as in (10).

Taking the conditional expectation w.r.t. $\mathcal{F}_{k-1}$ and using the tower property we get,

$$
\mathbb{E}\big[\bar{V}_t^k(\phi_t(s_t^k)) \mid \mathcal{F}_{k-1}\big] \leq \mathbb{E}\left[\sum_{t'=t}^{t+h-1} r(s_{t'}^k, a_{t'}^k) + \bar{V}_{t+h}^{k-1}(\phi_{t+h}(s_{t+h}^k)) \mid \mathcal{F}_{k-1}\right].
$$

Denote $s^k_{\phi,t} := \phi_t(s^k_t)$. Summing the above relation for all $n \in \{0\} \cup [\frac{H}{h} - 1]$, using linearity of expectation, and the fact $\bar{V}^k_{H+1}(\phi_{H+1}(s)) = 0$ for all s, k,

$$\sum_{n=0}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^k_{1+nh}(s^k_{\phi,1+nh}) \mid \mathcal{F}_{k-1}\big] \leq \mathbb{E}\left[\sum_{t=1}^{H} r(s^k_t, a^k_t) \mid \mathcal{F}_{k-1}\right] + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^{k-1}_{1+nh}(s^k_{\phi,1+nh}) \mid \mathcal{F}_{k-1}\big]$$

$$\iff \bar{V}^k_1(s^k_{\phi,1}) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^k_{1+nh}(s^k_{\phi,1+nh}) \mid \mathcal{F}_{k-1}\big] \leq \mathbb{E}\left[\sum_{t=1}^{H} r(s^k_t, a^k_t) \mid \mathcal{F}_{k-1}\right] + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^{k-1}_{1+nh}(s^k_{\phi,1+nh}) \mid \mathcal{F}_{k-1}\big]$$

$$\iff \bar{V}^k_1(s^k_{\phi,1}) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^k_{1+nh}(s^k_{\phi,1+nh}) \mid \mathcal{F}_{k-1}\big] \leq V^{\pi_k}(s^k_1) + \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^{k-1}_{1+nh}(s^k_{\phi,1+nh}) \mid \mathcal{F}_{k-1}\big]$$

$$\iff \bar{V}^k_1(s^k_{\phi,1}) - V^{\pi_k}(s^k_1) \leq \sum_{n=1}^{\frac{H}{h}-1} \mathbb{E}\big[\bar{V}^{k-1}_{1+nh}(s^k_{\phi,1+nh}) - \bar{V}^k_{1+nh}(s^k_{\phi,1+nh}) \mid \mathcal{F}_{k-1}\big]$$

$$\iff \bar{V}^k_1(s^k_{\phi,1}) - V^{\pi_k}(s^k_1) \leq \sum_{k=1}^{K} \sum_{n=1}^{\frac{H}{h}-1} \sum_{s_\phi \in \mathcal{S}_\phi} \bar{V}^{k-1}_{nh+1}(s_\phi) - \mathbb{E}[\bar{V}^k_{nh+1}(s_\phi) \mid \mathcal{F}_{k-1}]$$

The second line holds by the fact s^k_1 is measurable w.r.t. $\mathcal{F}_{k-1}$, the third line holds since

$$V^{\pi_k}_1(s^k_1) = \mathbb{E}\left[\sum_{t=1}^{H} r(s^k_t, a^k_t) \mid \mathcal{F}_{k-1}\right].$$

The fifth line holds by Lemma 15 with $\bar{V}^k_t = g^k_t$ for $t = nh + 1$. Furthermore, we set $\tilde{\mathcal{S}}$ of Lemma 15 to be $\mathcal{S}_\phi$. See that the update of $\bar{V}^k_t$ occurs only at the visited state $s^k_{\phi,t} = \phi(s^k_t)$ of the abstracted state space. Furthermore, the update rule uses $\bar{V}^{k-1}_{\phi,t+1}$, i.e., it is measurable w.r.t. to $\mathcal{F}_{k-1}$, and it is valid to apply the lemma. □

Theorem 7 (Performance of h-RTDP-AA)**.** *Let $\epsilon, \delta > 0$. The following holds for h-RTDP-AA:*

1. *With probability* $1 - \delta$*, for all* $K > 0$*,* $\mathrm{Regret}(K) \leq \frac{9 S_\phi H(H-h)}{h} \ln(3/\delta) + \frac{H \epsilon_A}{h} K.$
2. *Let* $\Delta_A = H\epsilon_A$*. Then,* $\Pr\left\{\exists \epsilon > 0 \ : \ N_\epsilon^{\frac{\Delta_A}{h}} \geq \frac{9 S_\phi H(H-h)\ln(3/\delta)}{h\epsilon}\right\} \leq \delta.$

Before supplying with the proof observe the following remark.

Proof. We start by proving **claim (1)**. The following bounds on the regret hold.

$$\begin{aligned}
\mathrm{Regret}(K) &:= \sum_{k=1}^{K} V^*_1(s^k_1) - V^{\pi_k}_1(s^k_1) \\
&\leq \sum_{k=1}^{K} \max_{s \in \Phi_1(s^k_1)} V^*_1(s) - V^{\pi_k}_1(s^k_1) \\
&\leq \sum_{k=1}^{K} \bar{V}^k_1(\phi_1(s^k_1)) - V^{\pi_k}_1(s^k_1) + \epsilon_A \frac{H}{h} \\
&\leq \epsilon_A \frac{H}{h} K + \sum_{k=1}^{K} \sum_{n=1}^{\frac{H}{h}-1} \sum_{s_\phi \in \mathcal{S}_\phi} \bar{V}^{k-1}_{nh+1}(s_\phi) - \mathbb{E}[\bar{V}^k_{nh+1}(s_\phi) \mid \mathcal{F}_{k-1}] \qquad (34)
\end{aligned}$$

The second relation holds the definition of the $\max$ operator and since $s^k_1 \in \phi(s^k_1)$ (by definition we have that $s \in \Phi_t(s)$, as $\phi_t(s) = \phi_t(s)$ for any t). The third relation holds by the approximate optimism of the value function (Lemma 13), and the forth relation is by Lemma 14.

We now observe the regret is a regret of a Decreasing Bounded Process. Let

$$X_k := \sum_{n=1}^{\frac{H}{h}-1} \sum_{s_\phi \in \mathcal{S}_\phi} \bar{V}^k_{nh+1}(s_\phi), \tag{35}$$

and observe that $\{X_k\}_{g \geq 0}$ is a Decreasing Bounded Process.

1. It is decreasing since for all $s_\phi \in \mathcal{S}_\phi, t$ $\bar{V}^k_t(s_\phi) \leq \bar{V}^{k-1}_t(s_\phi)$ by Lemma 13. Thus, their sum is also decreasing.
2. It is bounded since for all $s \in \mathcal{S}_\phi, t$ $\bar{V}^k_t(s_\phi) \geq 0$ by Lemma 13. Thus, the sum is bounded from below by 0.

See that the initial value can be bounded as follows,

$$X_0 = \sum_{n=1}^{\frac{H}{h}-1} \sum_{s_\phi \in \mathcal{S}_\phi} \bar{V}^0_{nh+1}(s_\phi) \leq \sum_{n=1}^{\frac{H}{h}-1} \sum_{s_\phi \in \mathcal{S}_\phi} H = \frac{S_\phi H(H-h)}{h}.$$

Using linearity of expectation and the definition (14) we observe that (34) can be written,

$$\text{Regret}(K) \leq (34) = \epsilon_A \frac{H}{h} K + \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}],$$

which is regret of A Bounded Decreasing Process. Applying the bound on the regret of a DRP, Theorem 1, we conclude the proof of the first claim.

We now prove **claim (2)** using the proving technique at Theorem 4. Denote $\Delta_A = H\epsilon_A$. The following relations hold for all $\epsilon > 0$.

$$\begin{aligned}
&\mathbb{1}\{\bar{V}^k_1(\phi_1(s^k_1)) - V^{\pi_k}_1(s^k_1) \geq \epsilon\}\epsilon \\
&\leq \mathbb{1}\{\bar{V}^k_1(\phi_1(s^k_1)) - V^{\pi_k}_1(s^k_1) \geq \epsilon\}\left(\bar{V}^k_1(\phi_1(s^k_1)) - V^{\pi_k}_1(s^k_1)\right) \\
&\leq \mathbb{1}\{\bar{V}^k_1(\phi_1(s^k_1)) - V^{\pi_k}_1(s^k_1) \geq \epsilon\}\left(\sum_{n=1}^{\frac{H}{h}-1} \sum_{s_\phi \in \mathcal{S}_\phi} \bar{V}^{k-1}_{nh+1}(s_\phi) - \mathbb{E}[\bar{V}^k_{nh+1}(s_\phi) \mid \mathcal{F}_{k-1}]\right) \\
&= \mathbb{1}\{\bar{V}^k_1(s^k_1) - V^{\pi_k}_1(s^k_1) \geq \epsilon\}(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]).
\end{aligned} \tag{36}$$

The first relation holds by the indicator function and the second relation holds by Lemma 14. The forth relation holds by the definition of X_k (35) and linearity of expectation.

As we wish the final performance to be compared to V^* we use the the first claim of Lemma 13, by which for all s, k, $\bar{V}^k_1(\phi_1(s)) \geq V^*_1(s) - \frac{\Delta_A}{h}$. This implies that

$$\mathbb{1}\left\{V^*_1(s^k_1) - V^{\pi_k}_1(s^k_1) \geq \frac{\Delta_A}{h} + \epsilon\right\} \leq \mathbb{1}\{\bar{V}^k_1(\phi_1(s^k_1)) - V^{\pi_k}_1(s^k_1) \geq \epsilon\}. \tag{37}$$

Combining the above we get

$$\begin{aligned}
&\mathbb{1}\left\{V^*_1(s^k_1) - V^{\pi_k}_1(s^k_1) \geq \frac{\Delta_A}{h} + \epsilon\right\}\epsilon \\
&\leq \mathbb{1}\{\bar{V}^k_1(\phi_1(s^k_1)) - V^{\pi_k}_1(s^k_1) \geq \epsilon\}\epsilon \\
&\leq \mathbb{1}\{\bar{V}^k_1(\phi_1(s^k_1)) - V^{\pi_k}_1(s^k_1) \geq \epsilon\}(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]).
\end{aligned} \tag{38}$$

The first relation is by (37) and the second relation by (36).

Define $N_\epsilon(K) = \sum_{k=1}^{K} \mathbb{1}\{V^*_1(s^k_1) - V^{\pi_k}_1(s^k_1) \geq \frac{\Delta_A}{h} + \epsilon\}$ as the number of times $V^*_1(s^k_1) - V^{\pi_k}_1(s^k_1) \geq \frac{\Delta_A}{h} + \epsilon$ at the first K episodes. Summing the above inequality (38) for all $k \in [K]$ and denote we get that for all $\epsilon > 0$

$$
\begin{aligned}
N_\epsilon(K)\epsilon &= \sum_{k=1}^{K} \mathbb{1}\left\{ V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \geq \frac{\Delta_A}{h} + \epsilon \right\}\epsilon \\
&\leq \sum_{k=1}^{K} \mathbb{1}\left\{ \bar{V}_1^k(\phi_1(s_1^k)) - V_1^{\pi_k}(s_1^k) \geq \epsilon \right\}(X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}]) \\
&\leq \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}].
\end{aligned}
$$

The first relation holds by definition, the second by (38) and the third relation holds as $\{X_k\}_{k\geq 0}$ is a DBP (35) and, thus, $X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \geq 0$ a.s. . Thus, the following relation holds

$$
\left\{ \forall K > 0 : \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \leq \frac{9SH(H-h)}{h} \ln \frac{3}{\delta} \right\} \subseteq \left\{ \forall \epsilon > 0 : N_\epsilon(K)\epsilon \leq \frac{9SH(H-h)}{h} \ln \frac{3}{\delta} \right\},
$$

from which we get that for any $K > 0$

$$
\begin{aligned}
&\Pr\left(\forall \epsilon > 0 : N_\epsilon(K)\epsilon \leq \frac{9SH(H-h)}{h} \ln \frac{3}{\delta} \right) \\
&\geq \Pr\left(\forall K > 0 : \sum_{k=1}^{K} X_{k-1} - \mathbb{E}[X_k \mid \mathcal{F}_{k-1}] \leq \frac{9SH(Hh)}{h} \ln \frac{3}{\delta} \right) \geq 1 - \delta,
\end{aligned}
$$

and the third relation holds the bound on the regret of DBP, Theorem 1. Equivalently, for any $K > 0$,

$$
\Pr\left(\exists \epsilon > 0 : N_\epsilon(K)\epsilon \geq \frac{9SH(H-h)}{h} \ln \frac{3}{\delta} \right) \leq \delta. \tag{39}
$$

Applying the Monotone Convergence Theorem as in the proof of Theorem 4 we conclude the proof.

□

14 Useful Lemmas

The following lemma is a generalization of Lemma 34 in [15].

Lemma 15 (On Trajectory Regret to Uniform Regret). *For any $t \in [H]$, let $\left\{s_t^k, \mathcal{F}_k\right\}_{k\geq 0}$ be a random process where $\left\{s_t^k\right\}_{k\geq 0}$ is adapted to the filtration $\{\mathcal{F}_k\}_{k\geq 0}$ and $s_t^k \in \tilde{\mathcal{S}}$ where $\tilde{\mathcal{S}}$ is a finite set of all possible realizations of s_t^k with cardinally $\tilde{S} := |\tilde{\mathcal{S}}|$. Let $g_t^k \in \mathbb{R}^{\tilde{S}}$ and denoting the $s \in \tilde{\mathcal{S}}$ entry of the vector as $g_t^k(s)$. Furthermore, let $g_t^k(s)$ be updated only at the state s_t^k by an update rule which is $\mathcal{F}_{k-1}$ measurable, i.e.,*

$$g_t^k(s) = \begin{cases} f_t^{k-1}(s), \text{if } s = s_t^k, \\ g_t^{k-1}(s), \text{o.w.}. \end{cases}$$

Where $f_t^{k-1}(s)$ is an update rule $\mathcal{F}_{k-1}$ measurable. Then,

$$\sum_{k=1}^{K} \mathbb{E}[g_t^{k-1}(s_t^k) - g_t^k(s_t^k) \mid \mathcal{F}_{k-1}] = \sum_{k=1}^{K} \sum_{s\in\tilde{\mathcal{S}}} g_t^{k-1}(s) - \mathbb{E}[g_t^k(s) \mid \mathcal{F}_{k-1}]$$

Proof. The following relations hold.

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{t=1}^{H} \mathbb{E}[g_t^{k-1}(s_t^k) - g_t^k(s_t^k) \mid \mathcal{F}_{k-1}] \\
&= \sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{s\in\tilde{\mathcal{S}}} \mathbb{E}[\mathbb{1}\left\{s = s_t^k\right\} g_t^{k-1}(s) - \mathbb{1}\left\{s = s_t^k\right\} g_t^k(s) \mid \mathcal{F}_{k-1}] \\
&\overset{(1)}{=} \sum_{k=1}^{K}\sum_{t=1}^{H}\sum_{s\in\tilde{\mathcal{S}}} \mathbb{E}[\mathbb{1}\left\{s = s_t^k\right\} g_t^{k-1}(s) - \mathbb{1}\left\{s = s_t^k\right\} f_t^{k-1}(s) \mid \mathcal{F}_{k-1}] \\
&\overset{(2)}{=} \sum_{t=1}^{H}\sum_{s\in\tilde{\mathcal{S}}}\sum_{k=1}^{K} \mathbb{E}[\mathbb{1}\left\{s = s_t^k\right\} g_t^{k-1}(s) + \mathbb{1}\left\{s \neq s_t^k\right\} g_t^{k-1}(s) \mid \mathcal{F}_{k-1}] \\
&\qquad\qquad - \mathbb{E}[\mathbb{1}\left\{s = s_t^k\right\} f_t^{k-1}(s) + \mathbb{1}\left\{s \neq s_t^k\right\} g_t^{k-1}(s) \mid \mathcal{F}_{k-1}] \\
&\overset{(3)}{=} \sum_{t=1}^{H}\sum_{s\in\tilde{\mathcal{S}}}\sum_{k=1}^{K} g_t^{k-1}(s) - \mathbb{E}[\mathbb{1}\left\{s = s_t^k\right\} f_t^{k-1}(s) + \mathbb{1}\left\{s \neq s_t^k\right\} g_t^{k-1}(s) \mid \mathcal{F}_{k-1}] \\
&\overset{(4)}{=} \sum_{t=1}^{H}\sum_{s\in\tilde{\mathcal{S}}}\sum_{k=1}^{K} g_t^{k-1}(s) - \mathbb{E}[g_t^k(s) \mid \mathcal{F}_{k-1}].
\end{aligned} \tag{40}$$

Relation (1) holds since for $s = s_t^k$ the vector g_k^t is updated according by f^{k-1}. Relation (2) holds by adding and subtracting $\mathbb{1}\left\{s \neq s_t^k\right\} g_t^{k-1}(s)$ while using the linearity of expectation. (3) holds since for any event $\mathbb{1}\{A\} + \mathbb{1}\{A^c\} = 1$ and since g_t^{k-1} is $\mathcal{F}_{k-1}$ measurable. (4) holds by the definition of the update rule,

$$\begin{aligned}
&\mathbb{E}[\mathbb{1}\left\{s = s_t^k\right\} f_t^{k-1}(s) + \mathbb{1}\left\{s \neq s_t^k\right\} g_t^{k-1}(s) \mid \mathcal{F}_{k-1}] \\
&= \mathbb{E}[\mathbb{1}\left\{s = s_t^k\right\} \mid \mathcal{F}_{k-1}] f_t^{k-1}(s) + \mathbb{E}[\mathbb{1}\left\{s \neq s_t^k\right\} \mid \mathcal{F}_{k-1}] g_t^{k-1}(s) \\
&= \Pr(s_t^k = s \mid \mathcal{F}_{k-1}) f_t^{k-1}(s) + \Pr(s_t^k \neq s \mid \mathcal{F}_{k-1}) g_t^{k-1}(s) = \mathbb{E}[g_t^k(s) \mid \mathcal{F}_{k-1}].
\end{aligned}$$

Where we used that $g_t^{k-1}(s)$ is $\mathcal{F}_{k-1}$ measurable and the assumption that $f_t^{k-1}(s)$ is $\mathcal{F}_{k-1}$ measurable in the first relation. □

The following lemma is a variant of a well known error propagation analysis in case of an approximate model.

Lemma 16 (Model Error Propagation)**.** *Let* $\|(P(\cdot \mid s,a) - \hat{P}(\cdot \mid s,a))\| \le \epsilon_P$ *for any* s, a. *Then, for any policy* π*,*

$$\forall s_1 \in \mathcal{S},\ \sum_{s_n} \Big| P^\pi(s_n \mid s_1) - \hat{P}^\pi(s_n \mid s_1) \Big| \le n\epsilon_P$$

Proof. We prove the claim by induction. For the base case $n = 1$ we get that for any $s_1 \in \mathcal{S}$

$$\begin{aligned}
&\sum_{s_2} \Big| P^\pi(s_2 \mid s_1) - \hat{P}^\pi(s_2 \mid s_1) \Big| \\
&= \sum_{s_2} \Big| \sum_a \pi(a \mid s_1) \Big(P(s_2 \mid s_1, a) - \hat{P}^\pi(s_2 \mid s_1, a) \Big) \Big| \\
&\le \sum_a \pi(a \mid s_1) \sum_{s_2} \Big| P(s_2 \mid s_1, a) - \hat{P}^\pi(s_2 \mid s_1, a) \Big| \\
&= \sum_a \pi(a \mid s_1) \| P(\cdot \mid s_1, a) - P(\cdot \mid s_1, a) \|_1 \le \epsilon_P.
\end{aligned}$$

Assume the induction step, i.e., assume the claim holds for $k = n - 1$. We now prove the induction step, i.e., for $k = n$

$$\begin{aligned}
&\sum_{s_n} \Big| P^\pi(s_n \mid s_1) - \hat{P}^\pi(s_n \mid s_1) \Big| \\
&= \sum_{s_n} \Big| \sum_{s_2} P^\pi(s_n \mid s_2) P^\pi(s_2 \mid s_1) - \hat{P}^\pi(s_n \mid s_2) \hat{P}^\pi(s_2 \mid s_1) \Big| \\
&\le \sum_{s_n} \sum_{s_2} \Big| P^\pi(s_n \mid s_2) P^\pi(s_2 \mid s_1) - \hat{P}^\pi(s_n \mid s_2) \hat{P}^\pi(s_2 \mid s_1) \Big| \\
&\le \sum_{s_n} \sum_{s_2} \Big| P^\pi(s_n \mid s_2) P^\pi(s_2 \mid s_1) - \hat{P}^\pi(s_n \mid s_2) P^\pi(s_2 \mid s_1) \Big| \\
&\qquad + \Big| \hat{P}^\pi(s_n \mid s_2) \hat{P}^\pi(s_2 \mid s_1) - \hat{P}^\pi(s_n \mid s_2) P^\pi(s_2 \mid s_1) \Big| \\
&\le \sum_{s_n} \sum_{s_2} P^\pi(s_2 \mid s_1) \Big| P^\pi(s_n \mid s_2) - \hat{P}^\pi(s_n \mid s_2) \Big| \\
&\qquad + \hat{P}^\pi(s_n \mid s_2) \Big| \hat{P}^\pi(s_2 \mid s_1) - P^\pi(s_2 \mid s_1) \Big| \\
&\le \underbrace{\sum_{s_2} P^\pi(s_2 \mid s_1)}_{=1} \left(\max_{s_2'} \sum_{s_n} \Big| P^\pi(s_n \mid s_2') - \hat{P}^\pi(s_n \mid s_2') \Big| \right) \\
&\qquad + \sum_{s_2} \underbrace{\left(\sum_{s_n} \hat{P}^\pi(s_n \mid s_2) \right)}_{=1} \Big| \hat{P}^\pi(s_2 \mid s_1) - P^\pi(s_2 \mid s_1) \Big| \\
&= \max_{s_2} \sum_{s_n} \Big| P^\pi(s_n \mid s_2) - \hat{P}^\pi(s_n \mid s_2) \Big| + \sum_{s_2} \Big| \hat{P}^\pi(s_2 \mid s_1) - P^\pi(s_2 \mid s_1) \Big|.
\end{aligned}$$

By the induction hypothesis and the base case,

$$\begin{aligned}
&\max_{s_2'} \sum_{s_n} \Big| P^\pi(s_n \mid s_2') - \hat{P}^\pi(s_n \mid s_2') \Big| \le \epsilon(n-1) \\
&\sum_{s_2} \Big| \hat{P}^\pi(s_2 \mid s_1) - P^\pi(s_2 \mid s_1) \Big| \le \epsilon_P,
\end{aligned}$$

from which we prove the induction step,

$$\forall s_1 \in \mathcal{S},\ \|P^\pi(\cdot \mid s_1)_1 - \hat{P}^\pi(\cdot \mid s_1)\| = \sum_{s_n} \Big| P^\pi(s_n \mid s_1) - \hat{P}^\pi(s_n \mid s_1) \Big| \le n\epsilon_P.$$

□

Lemma 17. *Let $V_t^*(s), \hat{V}_t^*(s)$ be the optimal values on the MDP $\mathcal{M}, \hat{\mathcal{M}}$, respectively, and let $V_t^\pi(s), \hat{V}_t^\pi(s)$ be the value of a fixed policy π on the MDP $\mathcal{M}, \hat{\mathcal{M}}$, respectively. Then,*

$$i)\ ||V_1^* - \hat{V}_1^*||_\infty \le \frac{H(H-1)}{2}\epsilon_P,$$
$$ii)\ \forall \pi, \|V_1^\pi - \hat{V}_1^\pi\|_\infty \le \frac{H(H-1)}{2}\epsilon_P.$$

Proof. Both claims follow standard techniques based on the Simulation Lemma [20, 30].

(i) Let $\Delta_t(s) := \hat{V}_t^*(s) - V_t^*(s), \Delta_t = \max_s |\Delta_t(s)|$. For $t = H$ we have that for all s

$$\begin{aligned}\Delta_H(s) &= \max_a r(s,a) + \sum_{s'} \hat{P}(s' \mid s,a)\hat{V}_{H+1}^*(s') - \max_a r(s,a) + \sum_{s'} P(s' \mid s,a) V_{H+1}^*(s') \\ &= \max_a r(s,a) + \sum_{s'} \hat{P}(s' \mid s,a) \cdot 0 - \max_a r(s,a) + \sum_{s'} P(s' \mid s,a) \cdot 0 = 0, \end{aligned} \tag{41}$$

and the base case holds. Assume the claim holds for any $t \ge k+1$, we now prove it holds for $t = k$. The following relations hold for any s,

$$\begin{aligned}\Delta_t(s) &= \max_a r(s,a) + \sum_{s'} \hat{P}(s' \mid s,a)\hat{V}_{t+1}^*(s') - \max_a r(s,a) + \sum_{s'} P(s' \mid s,a) v_{t+1}^*(s') \\ &\le r(s,a^*) + \sum_{s'} \hat{P}(s' \mid s,a^*)\hat{V}_{t+1}^*(s') - r(s,a^*) + \sum_{s'} P(s' \mid s,a^*) V_{t+1}^*(s') \\ &= \sum_{s'} \hat{P}(s' \mid s,a^*)\hat{V}_{t+1}^*(s') - P(s' \mid s,a^*) V_{t+1}^*(s') \\ &\le \sum_{s'} \hat{P}(s' \mid s,a^*) \underbrace{\Big|\hat{V}_{t+1}^*(s') - V_{t+1}^*(s')\Big|}_{:=\Delta_{t+1}(s')} + \Big|P(s' \mid s,a^*) - \hat{P}(s' \mid s,a^*)\Big| V_{t+1}^*(s') \\ &\le \sum_{s'} \hat{P}(s' \mid s,a^*)|\Delta_{t+1}(s')| + (H-t)\epsilon_P \\ &\le \Delta_{t+1} \sum_{s'} \hat{P}(s' \mid s,a^*) + (H-t)\epsilon_P = \Delta_{t+1} + (H-t)\epsilon_P \end{aligned}$$

The second relation holds by choosing a^* to maximize the first term first. The forth relation by adding and subtracting $\hat{P}(s' \mid s,a^*)\hat{V}_{t+1}^*(s')$ and standard inequalities. The fifth relation by the fact $V_{t+1}^*(s) \le H - t$ and the assumption that for all s, a $\|P(\cdot \mid s,a) - \hat{P}(\cdot \mid s,a)\| \le \epsilon_P$. The sixth by the fact $\hat{P}(\cdot \mid s,a)$ is a probability distribution and thus sums to 1.

Lower bounding $\Delta_t(s)$ using similar technique with opposite inequalities yields,

$$\Delta_t(s) \ge -(\Delta_{t+1} + (H-t)\epsilon_P),$$

and thus,

$$|\Delta_t(s)| \le \Delta_{t+1} + (H-t)\epsilon_P.$$

As the above holds for any s it holds for the maximizer. Thus,

$$\Delta_t \le \Delta_{t+1} + (H-t)\epsilon_P.$$

Iterating on this relation while using $\Delta_H(s) = 0$ by (41),

$$||V_1^* - \hat{V}_1^*||_\infty = \Delta_1 \le \sum_{t=1}^{H} (H-t)\epsilon_P = \epsilon_P \sum_{t=1}^{H-1} t = \frac{H(H-1)}{2}\epsilon_P.$$

(ii) The proof of the second claim follows the same proof of the first claim, without while replacing the $\max$ operator with the a fixed policy π. $\square$

Lemma 18 (Total Contribution of Approximate Model Errors)**.** *Let* $d_n := -\frac{1}{2}(h-1)h\epsilon_P + (H-n)h\epsilon_P$. *Then,*

$$\sum_{n=0}^{\frac{H}{h}-1} d_{1+nh} = \frac{1}{2}H(H-1)\epsilon_P.$$

Proof. The following relations hold.

$$\begin{aligned}
\sum_{n=0}^{\frac{H}{h}-1} d_{1+nh} &= -\frac{1}{2}H(h-1)\epsilon_P + \sum_{n=0}^{\frac{H}{h}-1}(H-1-nh)h\epsilon_P \\
&= -\frac{1}{2}H(h-1)\epsilon_P + H(H-1)\epsilon_P - h^2\epsilon_P\sum_{n=0}^{\frac{H}{h}-1} n \\
&= -\frac{1}{2}H(h-1)\epsilon_P + H(H-1)\epsilon_P - \frac{1}{2}h^2\epsilon_P(\frac{H-h}{h})\frac{H}{h} \\
&= -\frac{1}{2}H(h-1)\epsilon_P + H(H-1)\epsilon_P - \frac{1}{2}H(H-h)\epsilon_P \\
&= -\frac{1}{2}H(H-1)\epsilon_P + H(H-1)\epsilon_P = \frac{1}{2}H(H-1)\epsilon_P.
\end{aligned}$$

$\square$

15 Approximate Dynamic Programming in Finite-Horizon MDPs

Algorithm 9 (Exact) h-DP

init: $\forall s \in \mathcal{S},\ \forall n \in [\frac{H}{h}],\ V_{nh+1}(s) = H - nh$
for $n = \frac{H}{h} - 1, \frac{H}{h} - 2, \dots, 1$ **do**
  for $s \in \mathcal{S}$ **do**
    $V_{nh+1}(s) = \left(T^h V_{(n+1)h+1}\right)(s)$
  end for
end for
return: $\{V_{nh+1}\}_{n=1}^{H/h}$

Algorithm 10 h-DP with Approximate Model

init: $\forall s \in \mathcal{S},\ \forall n \in [\frac{H}{h}],\ V_{nh+1}(s) = H - nh$
for $n = \frac{H}{h} - 1, \frac{H}{h} - 2, \dots, 1$ **do**
  for $s \in \mathcal{S}$ **do**
    $V_{nh+1}(s) = \left(\hat{T}^h V_{(n+1)h+1}\right)(s)$
  end for
end for
return: $\{V_{nh+1}\}_{n=1}^{H/h}$

In this section, we follow standard analysis [20, 30] and establish bounds on the performance of approximate DP algorithms which update by an h-step optimal Bellman operator (2). We abbreviate this class of algorithms by h-ADP. See that unlike previous analysis [20, 30], we focus on finite horizon MDPs, which is the setup in which h-RTDP is analyzed. The different approximation setting we analyze in this section corresponds to the ones anlayzed for h-RTDP: approximate model, approximate value update, and approximate state abstraction.

As a reminder and for the sake of completeness, we start by considering h-DP Algorithm 9, which is the exact, approximate-free, version of the following h-ADP algorithms. The algorithm uses backward induction and a h-step optimal Bellman operator T^h by which it outputs the values $\{V_{nh+1}\}_{n=2}^{\frac{H}{h}}$. Notice that it holds $\{V_{nh+1}\}_{n=2}^{\frac{H}{h}} = \left\{V^*_{nh+1}\right\}_{n=2}^{\frac{H}{h}}$ by standard arguments on the Backward Induction algorithm. Furthermore, T^h can be solved by Backward induction with the total computational complexity of $O(SAh)$ by using Backward Induction. Thus, the total computational complexity of h-DP is $O(SAH)$ similar to the one of standard DP, e.g., Backward Induction.

In terms of space complexity, h-DP stores in memory $O(S\frac{H}{h})$ value entries. Observe that an h-greedy policy (3) w.r.t. $\{V_{nh+1}\}_{n=2}^{\frac{H}{h}}$ is an optimal policy, as these values are the optimal values as previously observed. Ultimately, one would like using these values to act in the environment by the optimal policy. If one uses the Forward-Backward DP (Section 9) to calculate such an h-greedy policy, then an extra $O(hS_h)$ space should be used, which results in total $O(S\frac{H}{h} + hS_h)$ space complexity (as in h-RTDP) that decrease in h if S_h is not too big (see Remark 2). Furthermore, the computational complexity of such approach is $O(HhAS_hS_1)$ which increases in h.

In next sections, we consider approximate settings of h-DP and establish that an h-greedy policy (3) w.r.t. the output values $\{V_{nh+1}\}_{n=2}^{\frac{H}{h}}$ has an equivalent performance to the asymptotic policy by which h-RTDP acts.

15.1 h-ADP with an Approximate Model

In the case of an approximate model, we replace the Bellman operator T used in h-DP with $\hat{T}$, the Bellman operator of the approximate model $\hat{p}$ instead the true one p (we assume r is exactly known, which correspond to the assumption made in Section 6.1). This results in Algorithm 10. Similarly to Section 6.1, we assume $\|\hat{p}(\cdot \mid s,a) - p(\cdot \mid s,a)\|_{TV} \le \epsilon_P$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Furthermore, denote π^*_P as the optimal policy of the approximate MDP.

Equivalently to h-DP, Algorithm 10 returns the optimal values of the *approximate model* (Algorithm 10 can be interpreted as exact h-DP applied on the approximate model). Thus, the h-greedy policy w.r.t. to the outputs of Algorithm 10 $\{V_{nh+1}\}_{n=2}^{\frac{H}{h}}$ is the optimal policy of the approximate MDP, π^*_P. The performance of π^*_P is measured by relatively to the performance of the optimal policy, i.e., we wish to bound $\|V_1^* - V_1^{\pi^*_P}\|_\infty$. This term represents the performance gap between the optimal policy of the 'real' MDP to the performance of the optimal policy of the approximate MDP evaluated on the real MDP, and is bounded in the following proposition.

Proposition 19. *Assume for all* $(s,a) \in \mathcal{S} \times \mathcal{A} : \|\hat{p}(\cdot \mid s,a) - p(\cdot \mid s,a)\|_{TV} \le \epsilon_P$ *and let* π_P^* *be the optimal policy of the approximate MDP. Then,*

$$\|V_1^* - V_1^{\pi_P^*}\|_\infty \le H(H-1)\epsilon_P.$$

Proof. Let $\hat{V}^{\pi_P^*}$ be the optimal value on the approximate MDP. By using the triangle inequality, the first and second claim of Lemma 17 we conclude the proof,

$$\|V_1^* - V_1^{\pi_P^*}\|_\infty \le \|V_1^* - \hat{V}_1^{\pi_P^*}\|_\infty + \|\hat{V}_1^{\pi_P^*} - V_1^{\pi_P^*}\|_\infty \le H(H-1)\epsilon_P.$$

□

Algorithm 11 h-DP with Approximate Value Updates

```
init: ∀s ∈ S, ∀n ∈ [H/h], V_{nh+1}(s) = H − nh
for n = H/h − 1, H/h − 2, ..., 1 do
   for s ∈ S do
      V̄_t^k(s) = ε_V(s) + (T^h V_{t+h})(s)
   end for
end for
return: {V_{nh+1}}_{n=1}^{H/h}
```

Algorithm 12 h-DP with Approximate State abstraction

```
init: ∀s ∈ S, ∀n ∈ [H/h], V_{nh+1}(s) = H − nh
for n = H/h − 1, H/h − 2, ..., 1 do
   for s ∈ S do
      V̄_{nh+1}(φ(s)) = min{ (T^h V_{(n+1)h+1})(s), V̄_{nh+1}^k(φ(s)) }
   end for
end for
return: {V_{nh+1}}_{n=1}^{H/h}
```

15.2 h-DP with Approximate Value Updates

In the case of a approximate value updates Algorithm 9 is replaced by an value updates with added noise $\epsilon_V(s)$, by which Algorithm 11 is formulated. Similarly to the assumption used for h-RTDP with approximate value updates (see Section 6.2) we assume for all $s \in \mathcal{S}$, $|\epsilon_V(s)| \le \epsilon_V > 0$. The following proposition bounds the performance of an h-greedy policy w.r.t. the values output by Algorithm 11.

Proposition 20. *Assume for all* $s \in \mathcal{S},\ |\epsilon_V(s)| \le \epsilon_V$. *Let* π_V^* *be the* h*-greedy policy* (3) *w.r.t. output Algorithm 11. Then,*

$$\|V_1^* - V_1^{\pi_V^*}\|_\infty \le \frac{2H}{h}\epsilon_V.$$

Proof. Let $\left\{\hat{V}_{nh+1}^*\right\}_{n=1}^{H/h}$ denote the output of Algorithm 11. We establish two claims which are of similarity to the two claims of Lemma 17. Combining the two we prove the result.

(i) The following relations hold for all $s \in \mathcal{S}$ and $n \in \{0\} \cup [\frac{H}{h} - 1]$.

$$\begin{aligned}
&\Delta_{1+nh}(s) := \hat{V}^*_{1+nh}(s) - V^*_{1+nh}(s)\\
&= \epsilon_V(s) + T^h \hat{V}^*_{1+(n+1)h}(s) - T^h V^*_{1+(1+n)h}(s')\\
&= \epsilon_V(s) + \max_{a_0,\ldots,a_{h-1}} \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + \hat{V}^*_{t+h}(s_h) \mid s_0 = s\right] - \max_{a_0,\ldots,a_{h-1}} \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + V^*_{t+h}(s_h) \mid s_0 = s\right]
\end{aligned} \tag{42}$$

The second relation holds by the updating equation and the third relation by definition (2). Let $\{\hat{a}_0, \hat{a}_1, .., \hat{a}_{h-1}\}$ be the set of policies maximizes the second terms, then, by plugging this sequence to the third term we necessarily decrease it. Thus,

$$\begin{aligned}
(42) &\leq \epsilon_V(s) + \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + \hat{V}^*_{t+h}(s_h) \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right]\\
&\quad - \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + V^*_{t+h}(s_h) \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right]\\
&= \epsilon_V(s) + \mathbb{E}\left[\hat{V}^*_{t+h}(s_h) - V^*_{t+h}(s_h) \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right]\\
&= \epsilon_V(s) + \mathbb{E}\left[\Delta_{1+(n+1)h}(s) \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right] \leq \epsilon_V + \|\Delta_{1+(n+1)h}\|_\infty.
\end{aligned}$$

The second relation holds by linearity of expectation, the third relation by definition, and the forth by assumption on $\epsilon_V(s)$ and by the standard bounded $E[X] \leq \|X\|_\infty$.

Repeating the above arguments while choosing the sequence which maximizes the third term in (42) allows us to lower bound (42) as follows

$$(42) \geq -\epsilon_V - \|\Delta_{1+(n+1)h}\|_\infty,$$

and thus,

$$\|\Delta_{1+nh}\|_\infty \leq \epsilon_V + \|\Delta_{1+(n+1)h}\|_\infty.$$

Solving the recursion while using $\Delta_{H+1}(s) = 0$ for all $s \in \mathcal{S}$ we get

$$\|\Delta_1\|_\infty \leq \frac{H}{h} \epsilon_V. \tag{43}$$

(ii) The following relations hold for all $s \in \mathcal{S}$ and $n \in [\frac{H}{h}]$.

$$\begin{aligned}
&\Delta^{\pi^*_V}_{1+nh}(s) := \hat{V}^*_{1+nh}(s) - V^{\pi^*_V}_{1+nh}(s)\\
&= \epsilon_V(s) + \max_{a_0,..,a_{h-1}} \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + \hat{V}^*_{1+(n+1)h}(s_h) \mid s_0 = s\right]\\
&\quad - \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + V^{\pi^*_V}_{1+(n+1)h}(s_h) \mid s_0 = s, \pi^*_V\right].
\end{aligned}$$

By definition, the sequence which maximizes the second term is π^*_V as it is the h-greedy policy w.r.t. $\hat{V}^*$. Using the linearity of expectation we get

$$\begin{aligned}
\Delta^{\pi^*_V}_{1+nh}(s) &= \epsilon_V(s) + \mathbb{E}\left[\hat{V}^*_{1+(n+1)h}(s_h) - V^{\pi^*_V}_{1+(n+1)h}(s_h) \mid s_0 = s, \pi^*_V\right]\\
&= \epsilon_V(s) + \mathbb{E}\left[\Delta^{\pi^*_V}_{1+(n+1)h}(s_{1+(n+1)h}) \mid s_0 = s, \pi^*_V\right].
\end{aligned}$$

As for all s, $|\epsilon_V(s)| \leq \epsilon_V$, using the triangle inequality and $E[X] \leq \|X\|_\infty$ we get the following recursion,

$$\|\Delta^{\pi^*_V}_{1+nh}\|_\infty \leq \epsilon_V + \|\Delta^{\pi^*_V}_{1+(n+1)h}\|_\infty.$$

Using $\|\Delta_{1+H}^{\pi_V^*}\|_\infty = 0$ we arrive to its solution,

$$\|\Delta_1^{\pi_V^*}\|_\infty \leq \frac{H}{h}\epsilon_V. \tag{44}$$

which proves the second needed result.

Finally, using the triangle inequality and the two proven claims, (43) and (44), we conclude the proof.

$$\|V_1^* - V_1^{\pi_V^*}\|_\infty \leq \|V_1^* - \hat{V}_1\|_\infty + \|\hat{V}_1^* - V_1^{\pi_V^*}\|_\infty = \|\Delta_1\|_\infty + \|\Delta_1^{\pi_V^*}\|_\infty \leq 2\frac{H}{h}\epsilon_V.$$

□

15.3 h-DP with Approximate State Abstraction

When an approximate state abstraction $\{\phi_{1+nh}\}_{n=0}^{\frac{H}{h}-1}$ is given, Algorithm 9 can be replaced by an exact value update in the reduced state space $\mathcal{S}_\phi$, as given in Algorithm 12. This corresponds to updating a value $V \in \mathbb{R}^{S_\phi}$, instead a value $\mathbb{R}^S$. An obvious advantage of such an algorithm, relatively to h-DP, is its reduced space complexity, as it only needs to store $O(\frac{H}{h}S_\phi)$ value entries, instead of $O(\frac{H}{h}S)$ as h-DP.

Yet, as seen in Algorithm 12, its computational complexity remains $O(SAH)$ as it needs to uniformly update on the entire (non-abstracted) state space. Would have we being given a representative from each equivalence classes under ϕ_{1+nh} for every $n \in \{0\} \cup [\frac{H}{h}]$[8] we could suggest an alternative Backward Induction algorithm with computational complexity of $O(S_\phi AH)$. However, as we do not assume access to this knowledge, we are obliged to scan the entire state space, without further assumptions.

The following proposition bounds the performance of an h-greedy policy w.r.t. the values output by Algorithm 12. Similarly to the analysis of the corresponding h-RTDP algorithm (see Section 6.3), we assume $\{\phi_{1+nh}\}_{n=0}^{\frac{H}{h}-1}$ satisfy Assumption 1.

Proposition 21. *Let $\{\phi_{1+nh}\}_{n=0}^{\frac{H}{h}-1}$ satisfy Assumption 1. Let $\left\{\hat{V}_{nh+1}^*\right\}_{n=1}^{\frac{H}{h}}$ denote the output of Algorithm 12 and let π_A^* be the h-greedy policy w.r.t. these approximate values* (3). *Then,*

$$\|V_1^* - V_1^{\pi_A^*}\|_\infty \leq \frac{H}{h}\epsilon_A.$$

Proof. We establish two claims which are of similarily to the two claims of Lemma 17 and Proposition 20. Combining the two we prove the result.

(i) The following relations hold for any $s \in \mathcal{S}$.

$$\begin{aligned}
&\hat{V}_{1+nh}^*(\phi_{1+nh}(s)) - V_{1+nh}^*(s) \\
&= T^h\hat{V}_{\phi,1+(n+1)h}^*(s) - T^hV_{1+(1+n)h}^*(s) \\
&= \max_{a_0,..,a_{h-1}} \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + \hat{V}_{1+(n+1)h}^*(\phi_{1+(n+1)h}(s_h)) \mid s_0 = s\right] \\
&\quad - \max_{a_0,..,a_{h-1}} \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + V_{1+(n+1)h}^*(s_h) \mid s_0 = s\right]
\end{aligned} \tag{45}$$

The second and third relation holds by the updating rule of Algorithm 12. Let $\{\hat{a}_0, \hat{a}_1, .., \hat{a}_{h-1}\}$ be the set of policies which maximizes the first term. Then, by plugging this sequence to the second

[8] Differently put, if we interpret ϕ as clustering multiple states $s \in \mathcal{S}$ together, we would require a single representative from each such a cluster.

term we necessarily decrease it, and the following holds.

$$(45) \leq \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + \hat{V}^*_{1+(n+1)h}(\phi_{1+(n+1)h}(s_h)) \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right]$$
$$- \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + V^*_{1+(n+1)h}(s_h) \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right]$$
$$= \mathbb{E}\left[\hat{V}^*_{1+(n+1)h}(\phi_{1+(n+1)h}(s_h)) - V^*_{1+(n+1)h}(s_h) \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right] \tag{46}$$

Where the second relation holds by linearity of expectation. By Assumption 1 the following inequality holds,

$$(46) \leq \mathbb{E}\left[\hat{V}^*_{1+(n+1)h}(\phi_{1+(n+1)h}(s_h)) - \max_{\bar{s}_h \in \Phi_{1+(n+1)h}(s_h)} V^*_{1+(n+1)h}(\bar{s}_h) + \epsilon_A \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right]$$
$$= \epsilon_A + \mathbb{E}\left[\hat{V}^*_{1+(n+1)h}(\phi_{1+(n+1)h}(s_h)) - \max_{\bar{s}_h \in \Phi_{1+(n+1)h}(s_h)} V^*_{1+(n+1)h}(\bar{s}_h) \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right]$$
$$\leq \epsilon_A + \max_s \left|\hat{V}^*_{1+(n+1)h}(\phi_{1+(n+1)h}(s)) - \max_{\bar{s} \in \Phi_{1+(n+1)h}(s)} V^*_{1+(n+1)h}(\bar{s})\right| \tag{47}$$

By choosing the sequence of polices which maximizes the second term in (45) and repeating similar arguments to the above we arrive to the following relations.

$$(45) \geq \mathbb{E}\left[\hat{V}^*_{1+(n+1)h}(\phi_{1+(n+1)h}(s_h)) - V^*_{1+(n+1)h}(s_h) \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right]$$
$$\geq \mathbb{E}\left[\hat{V}^*_{1+(n+1)h}(\phi_{1+(n+1)h}(s_h)) - \max_{\bar{s}_h \in \Phi_{1+(n+1)h}(s_h)} V^*_{1+(n+1)h}(\bar{s}_h) \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right]$$
$$\geq -\mathbb{E}\left[\left|\hat{V}^*_{1+(n+1)h}(\phi_{1+(n+1)h}(s_h)) - \max_{\bar{s}_h \in \phi^{-1}_{1+(n+1)h}(s_h)} V^*_{1+(n+1)h}(\bar{s}_h)\right| \mid s_0 = s, \{a_{t'}\}_{t'=0}^{h-1} = \{\hat{a}_{t'}\}_{t'=0}^{h-1}\right]$$
$$\geq -\max_s \left|\hat{V}^*_{1+(n+1)h}(\phi_{1+(n+1)h}(s)) - \max_{\bar{s} \in \Phi_{1+(n+1)h}(s)} V^*_{1+(n+1)h}(\bar{s})\right| \tag{48}$$

Let $\Delta_{\phi,1+nh}(s) := \hat{V}^*_{1+nh}(\phi_{1+nh}(s)) - \max_{\bar{s} \in \Phi_{1+nh}(s)} V^*_{1+nh}(\bar{s})$. The following upper bound holds,

$$\begin{aligned}
\Delta_{\phi,1+nh}(s) &:= \hat{V}^*_{1+nh}(\phi_{1+nh}(s)) - \max_{\bar{s} \in \phi^{-1}_{1+nh}(s)} V^*_{1+nh}(\bar{s}) \\
&\leq \hat{V}^*_{1+nh}(\phi_{1+nh}(s)) - V^*_{1+nh}(s) \\
&\leq \max_s |\hat{V}^*_{1+(n+1)h}(\phi_{1+(n+1)h}(s)) - \max_{\bar{s} \in \Phi_{1+(n+1)h}(s)} V^*_{1+(n+1)h}(\bar{s})| + \epsilon_A \\
&= \|\Delta_{\phi,1+(n+1)h}\|_\infty + \epsilon_A.
\end{aligned}$$

where the third relation is by (47). Furthermore, the following lower bounds holds,

$$\begin{aligned}
\Delta_{\phi,1+nh}(s) &:= \hat{V}^*_{1+nh}(\phi_{1+nh}(s)) - \max_{\bar{s} \in \Phi_{1+nh}(s)} V^*_{1+nh}(\bar{s}) \\
&\geq \hat{V}^*_{1+nh}(\phi_{1+nh}(s)) - V^*_{1+nh}(s) - \epsilon_A \\
&\geq -\max_s \left|\hat{V}^*_{1+(n+1)h}(\phi_{1+(n+1)h}(s)) - \max_{\bar{s} \in \Phi_{1+(n+1)h}(s)} V^*_{1+(n+1)h}(\bar{s})\right| - \epsilon_A \\
&= -\|\Delta_{\phi,1+(n+1)h}\|_\infty - \epsilon_A,
\end{aligned}$$

where the second relation holds by Assumption 1 and the third by (48).

By the upper and lower bounds on $\Delta_{\phi,1+nh}(s)$ which holds for all s we conclude that

$$\|\Delta_{\phi,1+nh}\|_\infty \leq \|\Delta_{\phi,1+(n+1)h}\|_\infty + \epsilon_A.$$

Using $\|\Delta_{\phi,H+1}\|_\infty = 0$ we solve the recursion and conclude that

$$\|\Delta_{\phi,1}\|_\infty \leq \frac{H}{h}\epsilon_A. \tag{49}$$

(ii) The following relations hold based on similar arguments as in (42). Let $\Delta^{\pi^*_A}_{1+nh} := \max_s \hat{V}^*_{1+nh}(\phi_{1+nh}(s)) - V^{\pi^*_A}_{1+nh}(s)$. For all s the following relations hold.

$$\begin{aligned}
&\hat{V}^*_{1+nh}(\phi(s)) - V^{\pi^*_A}_{1+nh}(s) \\
&\leq \max_{a_0,..,a_{h-1}} \mathbb{E}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + \hat{V}^*_{1+(n+1)h}(\phi(s_h)) \mid s_0 = s\right] \\
&\quad - \mathbb{E}^{\pi^*_A}\left[\sum_{t'=0}^{h-1} r(s_{t'}, a_{t'}(s_{t'})) + V^{\pi^*_A}_{1+(n+1)h}(s_h) \mid s_0 = s\right],
\end{aligned} \tag{50}$$

the first relation holds by the updating rule which update by the (see Algorithm 12), and since $V^\pi_t = (T^\pi)^h V^\pi_{t+h}$, similarly to the optimal Bellman operator (2).

By definition, the sequence which maximizes the first term is π^*_A as it is the h-greedy policy w.r.t. $\hat{V}^*$. Using the linearity of expectation we get

$$\begin{aligned}
(50) &= \mathbb{E}^{\pi^*_A}\left[\hat{V}^*_{1+(n+1)h}(\phi(s_h)) - V^{\pi^*_A}_{1+(n+1)h}(s_h) \mid s_0 = s\right] \\
&\leq \max_s \hat{V}^*_{1+(n+1)h}(\phi(s)) - V^{\pi^*_A}_{1+(n+1)h}(s) := \Delta^{\pi^*_A}_{1+(n+1)h}.
\end{aligned} \tag{51}$$

Since (51) for all s it also holds for the maximum, i.e.,

$$\Delta^{\pi^*_A}_{1+nh} := \max_s \hat{V}^*_{1+nh}(\phi(s)) - V^{\pi^*_A}_{1+nh}(s) \leq \Delta^{\pi^*_A}_{1+(n+1)h}.$$

As $\Delta^{\pi^*_A}_{H+1} = 0$ and iterating on the above recursion we get,

$$\Delta^{\pi^*_A}_1 \leq 0. \tag{52}$$

We are now ready to prove the proposition. For any s the following holds,

$$V^*_1(s) - V^{\pi^*_A}_1(s) = \underbrace{V^*_1(s) - \hat{V}_1(\phi(s))}_{(A)} + \underbrace{\hat{V}_1(\phi(s)) - V^{\pi^*_A}_1(s)}_{B}.$$

By (49)

$$\begin{aligned}
(A) &\leq \max_{\bar{s}\in\Phi_1(s)} V^*_1(\bar{s}) - \hat{V}_1(\phi_1(s)) \\
&:= -\Delta_{\phi,1}(s) \leq \|\Delta_{\phi,1}\|_\infty \leq \frac{H}{h}\epsilon_A.
\end{aligned}$$

By (52),

$$\hat{V}_1(\phi_1(s)) - V^{\pi^*_A}_1(s) \leq \max_{\bar{s}}\Big(\hat{V}_1(\phi_1(\bar{s})) - V^{\pi^*_A}_1(\bar{s})\Big) = \Delta^{\pi^*_A}_1 \leq 0.$$

Lastly, combining the above and using $V^* \geq V^\pi$, we get that for all s

$$\begin{aligned}
&0 \leq V^*_1(s) - V^{\pi^*_A}_1(s) \leq \frac{H}{h}\epsilon_A. \\
&\to \|V^*_1 - V^{\pi^*_A}_1\|_\infty \leq \frac{H}{h}\epsilon_A.
\end{aligned}$$

□