NEAR-OPTIMAL REINFORCEMENT LEARNING WITH MULTI-STEP TRANSITION LOOKAHEAD

Anonymous authors
Paper under double-blind review

ABSTRACT

We study reinforcement learning (RL) with transition look-ahead, where the agent may observe which states would be visited upon playing any sequence of ℓ actions before deciding its course of action. Although look-ahead can substantially improve achievable performance, Pla et al. (2026a) showed that optimal planning with multi-step transition look-ahead is NP-hard. However, this hardness was established using a discount factor close to one. It was therefore unknown whether the problem remains hard for every discount factor, and whether near-optimal planning can nevertheless be performed efficiently. We resolve both questions. First, we show that for every fixed discount factor, exact planning remains NP-hard. Second, we introduce a randomized polynomial-time approximation scheme for every fixed look-ahead depth. Third, we extend our approach to account for unknown transitions. We empirically validate the soundness of our results on the wind-farm storage-control benchmark of Lu et al. (2025), showing that our approach, optimally accounting for ℓ-step look-ahead information, offers substantially better performance than existing algorithms.

1 INTRODUCTION

Reinforcement learning (RL) studies how an agent should act in a dynamic environment to maximize cumulative reward, accounting for both immediate gains and the long-term consequences of its decisions (Sutton & Barto, 2018). We consider stationary Markov decision processes (MDPs), whose reward function and transition kernel do not vary over time. RL with look-ahead extends this classical framework by providing the agent with predictive information before it acts. In the transition look-ahead setting, the agent observes, at each decision time, which states would be reached by every action sequence of length at most ℓ. Such information may arise from real-time forecasts, collaborative navigation systems, or high-fidelity simulators and world models. We refer to Pla et al. (2026a); Merlis (2024); Merlis et al. (2024); Lu et al. (2025) for further motivation and examples. Look-ahead can substantially improve achievable performance, but exploiting it requires policies and planning methods that account explicitly for this richer information structure. Standard RL algorithms do not provide such mechanisms directly, while simply discarding the additional information can be strictly suboptimal.

Related work. The idea of augmenting reinforcement learning with look-ahead information has recently received increasing attention. Merlis (2024) introduced a pseudo-polynomial algorithm for one-step transition look-ahead in the finite-horizon setting, while Merlis et al. (2024) studied general look-ahead horizons for *reward* look-ahead, in which rewards are revealed ahead of time, focusing on worst-case bounds on the ratio between the optimal values with and without look-ahead. Unfortunately, Pla et al. (2026a) ruled out the possibility of leveraging such extra information with reasonable computational cost: for perfect transition look-ahead, planning is polynomial-time solvable for $\ell = 1$ but becomes NP-hard for every $\ell \geq 2$. To sidestep this computational barrier, Lu et al. (2025) adopt a batched decision protocol: after observing an ℓ-step forecast, the agent commits to an entire block of actions before receiving a new forecast. This effectively reduces each block to a single decision and brings the problem back to a one-step look-ahead framework. However, it differs from the original rolling look-ahead setting, where the agent can revise its decisions as the prediction window shifts, and can therefore be strictly suboptimal. Whether near-optimal planning under genuine multi-step transition look-ahead can be performed efficiently thus remains open.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The hardness result also connects to the broader literature on the computational complexity of MDP planning. Classical planning in stationary MDPs is polynomial-time solvable under the discounted criterion, with foundational complexity results due to Papadimitriou (1987); subsequent work has further investigated the complexity of MDP planning under different objectives and horizons (Mundhenk et al., 2000; Littman et al., 2013; Balaji et al., 2018; Chen & Wang, 2017). Beyond these classical formulations, changes in the information structure can fundamentally alter computational complexity. For instance, Walsh et al. (2009) show that planning with delayed feedback can become NP-hard because of the exponential blow-up of the augmented state space, while partial observability leads to even stronger hardness results (Papadimitriou, 1987). Transition look-ahead exhibits a complementary phenomenon: providing additional information also induces a large augmented state space and makes exact planning intractable (Pla et al. (2026a)). This raises a natural algorithmic question: can such planning problems nevertheless be approximated efficiently without explicitly constructing the corresponding augmented MDP?

Two related paradigms provide useful starting points. First, Kearns et al. (2002) introduced sparse sampling, which computes near-optimal actions by exploring only a randomly sampled portion of the full look-ahead tree. Its complexity, however, is exponential in the effective horizon $(1-\gamma)^{-1}$. Second, sample-average approximation replaces intractable expectations with a fixed empirical sample and solves the resulting finite optimization problem (Kleywegt et al., 2002). Adapting this principle to transition look-ahead raises two challenges that have not been addressed: obtaining polynomial dependence on $(1-\gamma)^{-1}$ for fixed look-ahead depth ℓ, and ensuring that the precomputed solution remains accurate for every look-ahead window that may be observed at deployment.

Transition look-ahead can be viewed as prediction revealed before the agent acts. This connects our setting to the growing literature on algorithms with predictions. This perspective has recently been explored in several sequential decision-making settings. Li et al. (2024) design a learning-augmented controller for LQR with latent perturbations, where accurate predictions lead to near-optimal performance while robustness to prediction errors is retained. Lyu et al. (2026) study discounted MDPs equipped with predictions of the transition matrix and show that such predictions can reduce sample complexity, while Li et al. (2023) establish consistency–robustness trade-offs when the advice is provided in the form of predicted Q-values in non-stationary MDPs. Another line of work studies MDPs with exogenous information or dynamics. Pla et al. (2026b) consider discounted MDPs with i.i.d. exogenous contexts that are revealed before the agent acts and derive minimax PAC guarantees that exploit this structure, while Maran et al. (2026) study MDPs with markovian exogenous contexts and show that this structure can substantially improve learning guarantees.

Contribution. First, we refine the results of Pla et al. (2026a) by showing that the known NP-hardness does not rely on a large effective horizon: for every fixed rational discount factor $\gamma \in (0,1)$, exact planning remains NP-hard for $\ell \geq 2$ (Theorem 1). Second, for every fixed look-ahead depth ℓ, we give a randomized polynomial-time approximation scheme (RPTAS) that constructs, with high probability, a uniformly near-optimal policy (Theorem 2). We then extend our approach to unknown transitions with a regret learning algorithm (Theorem 3). Finally, on the wind-farm storage-control benchmark of Lu et al. (2025), our planning algorithm achieves substantially better performance than their planning algorithm.

2 Setting and objectives

2.1 Markov Decision Processes

We study finite tabular Markov decision processes (MDPs) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P(s' \mid s, a)$ denotes the probability of reaching state $s' \in \mathcal{S}$ after taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, and $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function. A (possibly randomized) stationary memoryless policy is a mapping $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the simplex over $\mathcal{A}$. Under such a policy, the interaction evolves as follows: at each time $t \in \mathbb{N}$, the system is in state $s_t \in \mathcal{S}$, the agent selects an action $a_t \sim \pi(\cdot \mid s_t)$, receives reward $r(s_t, a_t)$, and the next state is sampled according to $s_{t+1} \sim P(\cdot \mid s_t, a_t)$.

We consider the standard discounted-return objective. For a discount factor $\gamma \in (0,1)$ and an initial state $s \in \mathcal{S}$, the value of a policy π is $V^\pi(s) = \mathbb{E}^\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,|\, s_0 = s]$. The optimal discounted value function is defined by $V^*(s) = \sup_\pi V^\pi(s)$, $\forall s \in \mathcal{S}$, where the supremum is taken

over stationary memoryless policies. The optimal value function V^* is the unique solution to the Bellman optimality equations: $V^*(s) = \max_{a\in\mathcal{A}}\Big\{r(s,a)+\gamma\sum_{s'\in\mathcal{S}} P(s' \mid s,a)\, V^*(s')\Big\}, \forall s \in \mathcal{S}$.

2.2 Transition look-ahead

We now formalize the extra information provided by the look-ahead in terms of state observability and provide an augmented MDP construction that allows us to embed this problem into the standard evaluation framework introduced above. In the remainder of this paper, we focus on the perfect look-ahead case. The extension to noisy look-ahead is given in Section C.

2.2.1 Look-ahead and state observability

A convenient generative view of the transition kernel is the following. At each time t, for every state-action pair (s,a), the environment independently draws a potential successor $\Theta_t(s,a) \sim P(\cdot \mid s,a)$, collecting these draws defines a random transition table $\Theta_t : \mathcal{S}\times\mathcal{A} \to \mathcal{S}$, taking values in $\Omega := \mathcal{S}^{\mathcal{S}\times\mathcal{A}}$. Its distribution L is therefore

$$L(\theta) := \mathbb{P}\left(\Theta_t = \theta\right) = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} P\big(\theta(s,a) \mid s,a\big). \tag{1}$$

In a standard MDP, Θ_t is hidden from the agent and, after action A_t is selected, the realized transition is simply $S_{t+1} = \Theta_t(S_t, A_t)$. Transition look-ahead changes the information available before acting. At time t, the environment reveals some of the tables $\Theta_t, \Theta_{t+1}, \ldots$ to the agent.

Definition 1 (ℓ-step transition look-ahead)**.** *Let $\ell \geq 1$. At each decision time t, an agent with ℓ-step transition look-ahead observes, before choosing A_t, the window*

$$C_t^\ell := (\Theta_t, \Theta_{t+1}, \ldots, \Theta_{t+\ell-1}) \in \Omega^\ell. \tag{2}$$

This window encodes the complete depth-ℓ transition tree rooted at the current state. Indeed, starting from $s_0 = S_t$, any action sequence $a_0, \ldots, a_{k-1}$, with $k \leq \ell$, determines the trajectory

$$s_{j+1} = \Theta_{t+j}(s_j, a_j), \qquad j = 0, \ldots, k-1.$$

Thus, C_t^ℓ reveals the outcome of every action sequence of length at most ℓ before the first action is chosen. In particular, for $\ell = 1$, the agent knows the successor $\Theta_t(S_t, a)$ of every action $a \in \mathcal{A}$. After one transition, the oldest table is discarded and a fresh independent table is revealed, so that $C_{t+1}^\ell = (\Theta_{t+1}, \ldots, \Theta_{t+\ell-1}, \Theta_{t+\ell})$, with $\Theta_{t+\ell} \sim L$ independent of the previous tables. Hence, the pair (S_t, C_t^ℓ) evolves as a Markov process on $\mathcal{S}\times\Omega^\ell$, allowing the transition look-ahead problem to be treated as a standard discounted MDP on this augmented state space.

2.2.2 Policies and value functions

A stationary policy with ℓ-step transition look-ahead acts on the augmented state and is therefore a mapping $\pi : \mathcal{S}\times\Omega^\ell \to \Delta(\mathcal{A})$. Its value at an observed state-window pair $(s,c) \in \mathcal{S}\times\Omega^\ell$, with $c = (\theta_0, \ldots, \theta_{\ell-1})$, is $V_\ell^\pi(s,c) := \mathbb{E}^\pi\left[\sum_{t=0}^\infty \gamma^t r(S_t, A_t) \,\middle|\, S_0 = s,\ C_0^\ell = c\right]$.

We denote the optimal value by $V_\ell^\star(s,c) := \sup_\pi V_\ell^\pi(s,c)$. For any bounded function $V : \mathcal{S}\times\Omega^\ell \to \mathbb{R}$, define the Bellman optimality operator $(\mathcal{T}_\ell V)(s,\theta_0,\ldots,\theta_{\ell-1}) := \max_{a\in\mathcal{A}}\left\{r(s,a) + \gamma\mathbb{E}_{\Theta\sim L}\left[V\big(\theta_0(s,a),\theta_1,\ldots,\theta_{\ell-1},\Theta\big)\right]\right\}$. The operator $\mathcal{T}_\ell$ is a γ-contraction, so $V_\ell^\star$ is its unique fixed point. The corresponding optimal action-value function is $Q_\ell^\star(s,\theta_0,\ldots,\theta_{\ell-1},a) := r(s,a) + \gamma\mathbb{E}_{\Theta\sim L}\left[V_\ell^\star\big(\theta_0(s,a),\theta_1,\ldots,\theta_{\ell-1},\Theta\big)\right]$. An optimal policy is then obtained by choosing $\pi_\ell^\star(s,\theta_0,\ldots,\theta_{\ell-1}) \in \arg\max_{a\in\mathcal{A}} Q_\ell^\star(s,\theta_0,\ldots,\theta_{\ell-1},a)$.

3 Hardness of exact planning

Our first result strengthens the known hardness of multi-step transition look-ahead planning of Pla et al. (2026a) by showing that it persists for every fixed discount factor.

Theorem 1 (Fixed-discount hardness of transition look-ahead planning)**.** *Fix an integer $\ell \geq 2$ and a rational discount factor $\gamma \in (0, 1)$. Given a finite MDP $\mathcal{M}$, an initial state $s_0 \in \mathcal{S}$, and a rational threshold θ, deciding whether there exists a policy with perfect ℓ-step transition look-ahead such that $\mathbb{E}_{C\sim L^{\otimes \ell}}\left[V_\ell^\pi(s_0, C)\right] \geq \theta$ is NP-hard.*

Proof sketch. We sketch the proof for $\ell = 2$ and refer the reader to Section A.1 for a formal derivation. The proof is almost identical to that of Pla et al. (2026a). We reduce from the following expected-maximum subset problem of Mehta et al. (2020). Given independent, nonnegative, finite-support random variables $X_1, \ldots, X_n$, an integer k and a threshold $U \in \mathbb{Q}$, deciding whether there exists a set $S \subset [n]$ such that $|S| = k$ and $\mathbb{E}[\max_{i\in S} X_i] > U$ is NP-hard.

Given the input $(X_i)_{i\in[n]}, k, U$ we show that there exists an MDP with look-ahead $\ell = 2$ of size polynomial in the size of the input (in particular in n and the support size) such that finding the optimal policy is equivalent to finding the optimal set S^* that maximizes $\mathbb{E}[\max_{i\in S} X_i]$ over the set $\{S : S \subset [n], |S| = k\}$. The MDP is the following. At the root state s_0, the agent can either choose wait and stay at s_0, or choose go and go to s_1. At s_1, there are k actions $\mathsf{pick}_1, \ldots, \mathsf{pick}_k$, each transitioning independently and uniformly to one of the states $s_1^2, \ldots, s_n^2$ associated with the random variables. Consequently, the look-ahead window available at s_0 reveals a candidate set of states $\Sigma = \{q_1, \ldots, q_k\} \subset \{s_1^2, \ldots, s_n^2\}$ before the agent decides whether to leave s_0. Choosing wait reveals a fresh independent tuple while staying at s_0. If the agent chooses go, then at s_1 the look-ahead reveals the payoffs associated with Σ, allowing it to select the largest one. The rest of the MDP is chosen such that the payoffs associated with $s_1^2, \ldots, s_n^2$ have the same distribution as $X_1, \ldots, X_n$. If the optimal strategy is to wait until the best subset appears, then solving the MDP is the same as solving the expected-maximum subset problem.

However, if waiting does not give any reward and if the discount is low enough, it may be more rewarding to stop before seeing the optimal set. This is why the NP-hardness result in Pla et al. (2026a) requires the discount factor to be exponentially close to one so that the cost of waiting is negligible.Instead, in our proof, we set $T := \gamma^{\ell+1}U$ and assign the wait action a reward of $(1-\gamma)T$, so that waiting exactly preserves the threshold value T: $(1 - \gamma)T + \gamma T = T$. The rest of the proof follows Pla et al. (2026a).

□

4 Near-optimal planning

4.1 Algorithm

Despite this hardness of exact planning, near-optimal planning remains tractable: for every fixed look-ahead depth, we provide a randomized polynomial-time approximation scheme (RPTAS). Our algorithm proceeds in two phases. Offline, it samples a finite collection of transition tables and uses them to build a finite state space. Value iteration is then run on this sampled state space. Online, the agent combines the precomputed values with the actual look-ahead window it observes in order to select an action.

Offline phase. The offline phase constructs a finite proxy MDP by replacing the distribution L of transition tables with an empirical distribution. Given a dictionary size N, sample $\Theta^1, \ldots, \Theta^N \overset{\text{i.i.d.}}{\sim} L$ and define $\widehat{L}_N := \frac{1}{N}\sum_{j=1}^N \delta_{\Theta^j}$. Under $\widehat{L}_N$, every look-ahead window consists of tables from the dictionary and can therefore be identified with their indices. The proxy MDP consequently has the finite state space $\mathcal{D}_N := \mathcal{S} \times [N]^\ell$ where $(s, i_0, \ldots, i_{\ell-1})$ represents the augmented state $(s, \Theta^{i_0}, \ldots, \Theta^{i_{\ell-1}})$. From this state, action a leads to $\big(\Theta^{i_0}(s, a), i_1, \ldots, i_{\ell-1}, J\big),\ J \sim \mathrm{Unif}([N])$. Thus, the current table determines the next state, the remaining tables are shifted forward, and the new table is sampled uniformly from the dictionary.

Algorithm 1 approximately solves this proxy MDP by running K value-iteration steps on $\mathcal{D}_N$, starting from $V^0 = 0$, and returns $V^K : \mathcal{D}_N \to [0, V_{\max}],\ V_{\max} = \frac{1}{1-\gamma}$. If $V_N^\star$ denotes the optimal value function of the proxy MDP, contraction of its Bellman operator gives $\|V^K - V_N^\star\|_\infty \leq \gamma^K V_{\max}$. Hence, $K \geq \log(1/\eta)/(-\log\gamma)$ iterations suffice to achieve error at most $\eta V_{\max}$. Each iteration

takes $O(|\mathcal{S}||\mathcal{A}|N^{\ell+1})$ operations, and the resulting offline representation $D = (\Theta^1, \ldots, \Theta^N, V^K)$ is computed once, without enumerating the original space Ω^ℓ.

Online phase. At decision time, the agent observes its current state $s \in \mathcal{S}$ and an arbitrary look-ahead window $c = (\theta_0, \ldots, \theta_{\ell-1}) \in \Omega^\ell$. Since these tables need not belong to the sampled dictionary, the proxy value V^K cannot be evaluated directly at (s, c). The key observation is that, after ℓ transitions, every table in the current window has been shifted out and replaced by fresh tables, which are represented using the empirical distribution $\widehat{L}_N$. At that point, the resulting window belongs to the proxy state space $\mathcal{D}_N$, and V^K provides an appropriate terminal value.

Algorithm 2 therefore solves an ℓ-step planning problem with terminal value V^K. This problem is solved by backward induction because V^K is available only at the end of the horizon. For $m \in \{1, \ldots, \ell\}, u \in \mathcal{S}, i_0, \ldots, i_{m-1} \in [N]$, the quantity $U^{V^K}_{m,c}(u, i_0, \ldots, i_{m-1})$ represents the continuation value at step m, when the remaining tables from the observed window are $\theta_m, \ldots, \theta_{\ell-1}$ and the first m newly revealed tables are represented by $i_0, \ldots, i_{m-1}$. Formally, $U^{V^K}_{m,c}$ is defined recursively for $m \in \{1, \ldots, \ell\}$. The case $m = \ell$ is given by

$$U^{V^K}_{\ell,c}(u, i_0, \ldots, i_{\ell-1}) := V^K(u, i_0, \ldots, i_{\ell-1}), \tag{3}$$

for any $u \in \mathcal{S}, i_0, \ldots, i_{\ell-1} \in [N]$.

Then, for $m \in \{1, \ldots, \ell-1\}, u \in \mathcal{S}, i_0, \ldots, i_{m-1} \in [N]$, we define

$$U^{V^K}_{m,c}(u, i_0, \ldots, i_{m-1}) := \max_{a \in \mathcal{A}} \left\{ r(u, a) + \frac{\gamma}{N} \sum_{i_m=1}^{N} U^{V^K}_{m+1,c}\big(\theta_m(u, a), i_0, \ldots, i_m\big) \right\}. \tag{4}$$

Here, θ_m determines the next physical state, while the newly entering table is drawn uniformly from the dictionary. Importantly, each maximization is performed after conditioning on the tables available at that stage, so the recursion accounts for future adaptation rather than committing in advance to an open-loop action sequence. Finally, the score at state $s \in \mathcal{S}$, look-ahead window $c = (\theta_0, \ldots, \theta_{\ell-1}) \in \Omega^\ell$ and action $a \in \mathcal{A}$ is defined by

$$\widetilde{Q}_{V^K}(s, c, a) := r(s, a) + \frac{\gamma}{N} \sum_{i_0=1}^{N} U^{V^K}_{1,c}\big(\theta_0(s, a), i_0\big), \tag{5}$$

and the policy selects $\pi_D(s, c) \in \arg\max_{a \in \mathcal{A}} \widetilde{Q}_{V^K}(s, c, a)$. Only this first action is executed; at the next decision time, the procedure is repeated using the shifted look-ahead window.

Algorithm 1 RPTAS-offline

Input: MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$, look-ahead depth ℓ, dictionary size N, iterations K
Output: Policy representation D

1: Sample $\Theta^1, \ldots, \Theta^N \overset{\text{i.i.d.}}{\sim} L$
2: Set $\mathcal{D}_N \leftarrow \mathcal{S} \times [N]^\ell$
3: Initialize $V^0(s, i_0, \ldots, i_{\ell-1}) \leftarrow 0$ for all $(s, i_0, \ldots, i_{\ell-1}) \in \mathcal{D}_N$
4: **for** $k = 0, \ldots, K-1$ **do**
5: **for all** $(s, i_0, \ldots, i_{\ell-1}) \in \mathcal{D}_N$ **do**
6: $$V^{k+1}(s, i_0, \ldots, i_{\ell-1}) \leftarrow \max_{a \in \mathcal{A}} \{ r(s, a) + \frac{\gamma}{N} \sum_{j=1}^{N} V^k\big(\Theta^{i_0}(s, a), i_1, \ldots, i_{\ell-1}, j\big) \} \tag{6}$$
7: **end for**
8: **end for**
9: $D \leftarrow (\Theta^1, \ldots, \Theta^N, V^K)$
10: **return** D

Algorithm 2 RPTAS-online

Input: $D = (\Theta^1, \dots, \Theta^N, V^K)$, current state s, observed window $c = (\theta_0, \dots, \theta_{\ell-1})$
Output: Action $\pi_D(s, c)$

$U^{V^K}_{\ell,c}(u, i_0, \dots, i_{\ell-1}) \leftarrow V^K(u, i_0, \dots, i_{\ell-1})$ for all $(u, i_0, \dots, i_{\ell-1}) \in \mathcal{D}_N$
for $m = \ell - 1, \dots, 1$ **do**
  for all $(u, i_0, \dots, i_{m-1}) \in \mathcal{S} \times [N]^m$ **do**
    $U^{V^K}_{m,c}(u, i_0, \dots, i_{m-1}) \leftarrow \max_{a \in \mathcal{A}} \{ r(u, a) + \frac{\gamma}{N} \sum_{i_m=1}^{N} U^{V^K}_{m+1,c}\big(\theta_m(u, a), i_0, \dots, i_m\big) \}$
  end for
end for
for all $a \in \mathcal{A}$ **do**
  $\widetilde{Q}_{V^K}(s, c, a) \leftarrow r(s, a) + \frac{\gamma}{N} \sum_{i_0=1}^{N} U^{V^K}_{1,c}\big(\theta_0(s, a), i_0\big)$
end for
return $\pi_D(s, c) \leftarrow \arg\max_{a \in \mathcal{A}} \widetilde{Q}_{V^K}(s, c, a)$

4.2 Theoretical Guarantees

Theorem 2 (RPTAS with perfect ℓ-step look-ahead)**.** *Fix $\ell \geq 2$ and $\varepsilon, \delta \in (0, 1)$. For $N = \widetilde{O}(\frac{\ell |\mathcal{S}||\mathcal{A}|}{\varepsilon^2 (1-\gamma)^4})$ and $K = \widetilde{O}(\frac{1}{1-\gamma})$, Algorithms 1 and 2 compute a policy π such that*

$$\mathbb{P}\Big(V^{\pi}_{\ell}(s, c) \geq V^{\star}_{\ell}(s, c) - \varepsilon V_{\max}, \qquad \forall (s, c) \in \mathcal{S} \times \Omega^{\ell} \Big) \geq 1 - \delta. \tag{7}$$

For every fixed ℓ, computing and storing the policy, as well as selecting an action at each decision time, require time and memory polynomial in $|\mathcal{S}|$, $|\mathcal{A}|$, ε^{-1}, $\log(1/\delta)$, and $(1-\gamma)^{-1}$.

Proof sketch for $\ell = 2$. We separate the proof into three steps: (i) approximating the distribution of transition tables, (ii) solving the resulting empirical MDP on dictionary windows, and (iii) extending this solution to an arbitrary observed window.

(i) Let $\widehat{L}_N = N^{-1} \sum_{j=1}^{N} \delta_{\Theta^j}$, and let $\widehat{V}^{\star}$ and $\widehat{Q}^{\star}$ be the optimal value and action-value functions when L is replaced by $\widehat{L}_N$. In particular $\widehat{V}^{\star}$ is the fixed point of $\widehat{\mathcal{T}}_2$ defined by

$$(\widehat{\mathcal{T}}_2 V)(s, \theta_0, \theta_1) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \frac{\gamma}{N} \sum_{j=1}^{N} V\big(\theta_0(s, a), \theta_1, \Theta^j\big) \right\}. \tag{8}$$

Set $\eta := \varepsilon(1-\gamma)^2 V_{\max}/4$. The functions $\Theta \mapsto V^{\star}_2(u, \theta_1, \Theta)$ take values in $[0, V_{\max}]$, and there are at most $|\mathcal{S}||\Omega| = |\mathcal{S}|^{1+|\mathcal{S}||\mathcal{A}|}$ such functions. Hence, if $N \geq \frac{V^2_{\max}}{2\eta^2} \left(\log \frac{2}{\delta} + (1 + |\mathcal{S}||\mathcal{A}|) \log |\mathcal{S}| \right)$, Hoeffding's inequality and a union bound give, with probability at least $1 - \delta$, $\sup_{u, \theta_1} \left| \frac{1}{N} \sum_{j=1}^{N} V^{\star}_2(u, \theta_1, \Theta^j) - \mathbb{E}_{\Theta \sim L}[V^{\star}_2(u, \theta_1, \Theta)] \right| \leq \eta$. On this event, contraction of the Bellman operators yields $\|\widehat{V}^{\star} - V^{\star}_2\|_{\infty} \leq \frac{\gamma\eta}{1-\gamma}$, $\|\widehat{Q}^{\star} - Q^{\star}_2\|_{\infty} \leq \frac{\gamma\eta}{1-\gamma}$.

(ii) Now let us apply the Bellman optimality equation in the proxy MDP whose state space is $\mathcal{D}_N = \mathcal{S} \times [N]^2$, where (s, i_0, i_1) represents $(s, \Theta^{i_0}, \Theta^{i_1})$.

$$\begin{aligned} V^{*}_N(s, i_0, i_1) &= \max_{a \in \mathcal{A}} \left\{ r(s, a) + \frac{\gamma}{N} \sum_{j=1}^{N} \widehat{V}^{\star}\big(\Theta^{i_0}(s, a), \Theta^{i_1}, \Theta^j\big) \right\} \\ &= \widehat{\mathcal{T}}_2 \widehat{V}^{\star}(s, \Theta^{i_0}, \Theta^{i_1}) = \widehat{V}^{\star}(s, \Theta^{i_0}, \Theta^{i_1}) \end{aligned}$$

Thus, value iteration on $\mathcal{D}_N$ computes the empirical optimal value on every dictionary window. After K iterations, $\|V^K - V^{\star}_N\|_{\infty} \leq \gamma^K V_{\max}$.

(iii) Let $c = (\theta_0, \theta_1) \in \Omega^2$. Unrolling the empirical Bellman equation gives

$$\widehat{V}^\star(s, \theta_0, \theta_1) = \max_{a_0 \in \mathcal{A}} \Bigg\{ r(s, a_0) + \frac{\gamma}{N} \sum_{i_0=1}^{N} \max_{a_1 \in \mathcal{A}} \Bigg[r\big(\theta_0(s, a_0), a_1\big) + \frac{\gamma}{N} \sum_{i_1=1}^{N} \underbrace{\hat{V}^\star\Big(\theta_1\big(\theta_0(s, a_0), a_1\big), i_0, i_1\Big)}_{(II)} \Bigg] \Bigg\}. \tag{9}$$

In particular, (II) contains only dictionary tables. Hence, as proven at step 2, $(II) = V_N^\star\big(\theta_1(s_1, a_1), i_0, i_1\big)$. Therefore, for a fixed root action a_0, the expression inside the outer maximum is exactly the score computed by the online recursion. Thus, $\widetilde{Q}_{V_N^\star}(s, c, a_0) = \widehat{Q}^\star(s, c, a_0)$. See Appendix A.2 for the full proof and Appendix C.2 for its extension to noisy look-ahead. □

5 Regret minimization

We now consider the online setting in which the transition kernel P is initially unknown. Note that each revealed transition table provides one sample from every row $P(\cdot \mid s, a)$, regardless of the chosen action. Learning P therefore requires no exploration: we can repeatedly estimate the model and apply our RPTAS to it. Following the γ-regret of Liu & Su (2021), we apply this criterion to the augmented state $X_t = (S_t, C_t)$:

$$\mathrm{Reg}(T) := \sum_{t=0}^{T-1} \big[(1-\gamma) V_\ell^\star(X_t) - r_t\big], \qquad r_t = r(S_t, A_t). \tag{10}$$

The factor $1 - \gamma$ puts the optimal discounted value on the scale of a one-step reward, so the regret compares this benchmark at each encountered state with the rewards actually collected. To control this regret, RPTAS-Learn periodically rebuilds the dictionary and value array used for planning, using an empirical transition model. Updates occur at times $M \in \{1, 2, 4, \ldots\}$. At each update, it uses all tables preceding the current window to form

$$\widehat{P}_M(s' \mid s, a) := \frac{1}{M} \sum_{j=0}^{M-1} \mathbf{1}\{\Theta_j(s, a) = s'\}, \qquad \widehat{L}_M := \bigotimes_{s,a} \widehat{P}_M(\cdot \mid s, a). \tag{11}$$

This includes every state–action row, whether visited or not. The ℓ tables in the current window are used directly for action selection. We then run Algorithm 1 on $\widehat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \widehat{P}_M, r)$: draw N_M independent tables from $\widehat{L}_M$ and perform K_M ordinary value-iteration steps from zero. The resulting dictionary and value array are held fixed until the next update. At each decision, Algorithm 2 combines them with the current window to select an action. Both planning routines are unchanged. Updating only at powers of two requires $O(\log(T+1))$ planner recomputations while keeping the active sample size within a factor of two of the elapsed time.

Algorithm 3 RPTAS-Learn

Input: T, γ, ℓ, r, parameters (N_M, K_M)

Observe X_0 and execute a fixed action $A_0 \in \mathcal{A}$
for $t = 1, \ldots, T-1$ **do**
  Observe $X_t = (S_t, C_t^\ell)$
  if $t \in \{2^q : q \in \mathbb{N}_0\}$ **then**
    $M \leftarrow t$
    Compute $\widehat{P}_M$ using equation 11
    $D \leftarrow$ RPTAS-offline$(\widehat{\mathcal{M}}, \ell, N_M, K_M)$
  end if
  $A_t \leftarrow$ RPTAS-online(D, S_t, C_t^ℓ)
  Execute A_t and receive r_t
end for

We use a fixed rule to break ties, so each planner defines a deterministic stationary policy on the augmented state space.

Theorem 3 (Regret of RPTAS-LEARN). *Fix $\ell \geq 1$. At every update, choose $N_M = \left\lceil \frac{M}{(1-\gamma)^3} \right\rceil$ and $K_M = \left\lceil \frac{1}{1-\gamma} \log \frac{M^2}{1-\gamma} \right\rceil$. For every $T \geq 1$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathrm{Reg}(T) \leq \widetilde{O}_\ell \left(\sqrt{\frac{|\mathcal{S}||\mathcal{A}|T}{1-\gamma}} + \frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^2} \right). \tag{12}$$

Here $\widetilde{O}_\ell$ hides logarithmic factors in $|\mathcal{S}|, |\mathcal{A}|, T, 1/\delta$ and factors depending on ℓ. An update based on M tables costs $O\big(M|\mathcal{S}||\mathcal{A}| + N_M|\mathcal{S}||\mathcal{A}| + K_M|\mathcal{S}|(|\mathcal{A}|+1)N_M^\ell\big)$, and each subsequent action query costs $O\big(\ell|\mathcal{S}|(|\mathcal{A}|+1)N_M^\ell\big)$. Thus, for fixed ℓ, both costs are polynomial in $|\mathcal{S}|, |\mathcal{A}|, T$ and $(1-\gamma)^{-1}$.

Proof sketch. At each deterministic update time M, the observed tables provide M independent samples from every state–action row, exactly as under uniform sampling from a generative model. We therefore buid upon the empirical-model comparison and absorption argument of Azar et al. (2012), combining Bernstein concentration with a discounted variance bound. Applying this strategy to transition look-ahead requires controlling the product empirical law and the dictionary approximation; obtaining the γ-regret of Liu & Su (2021) additionally requires relating policy values to realised rewards.

The first difficulty is that $\widehat{L}_M$ is a product of empirical row distributions, rather than the empirical distribution of complete tables. By independently rematching the samples within each row, we obtain a Bernstein bound for its error on any fixed function, with variance under L (Lemma 9). Applied to continuations of $V_\ell^\star$, this bound combines with a discounted resolvent variance bound to control the propagation of model errors under each deployed policy (Lemma 11). To handle the dependence of $\widehat{V}_M^\star$ on the samples, we decompose $\widehat{V}_M^\star = V_\ell^\star + (\widehat{V}_M^\star - V_\ell^\star)$., control the second part uniformly using a product-law KL bound (Lemma 10), and absorb it in the Bellman comparison once $M \geq M_0 = \widetilde{O}(|\mathcal{S}|^2|\mathcal{A}|/(1-\gamma)^2)$.

Conditionally on the observed tables, the dictionary is freshly sampled from $\widehat{L}_M$. The known-model RPTAS analysis therefore controls its planning error, including at windows outside the dictionary. Together, these arguments bound the normalized loss of each deployed policy by $\widetilde{O}_\ell\Big(\sqrt{|\mathcal{S}||\mathcal{A}|/[M(1-\gamma)]} + |\mathcal{S}||\mathcal{A}|/[M(1-\gamma)] + M^{-2}\Big)$. Finally, the active policy is fixed between each lazy update. Its Bellman equation makes the value-to-reward differences telescope up to epoch boundaries; a variance-sensitive Freedman bound controls the remaining martingale. Summing the policy losses over epochs and bounding the initial $M < M_0$ steps trivially gives the claim. The full proof is given in Appendix A.3. □

6 NUMERICAL STUDIES

We evaluate RPTAS on the wind-farm storage-control benchmark of Lu et al. (2025). The wind farm incurs penalties when production falls short of its scheduled production, while a capacity-constrained battery can store surplus energy or compensate for later shortages. Choosing when to charge or discharge therefore depends critically on forecasts of future prices and production mismatches.

Our theory is presented for perfect look-ahead, whereas the benchmark provides noisy forecasts. As shown in Section C.2, RPTAS extends to this setting by replacing each revealed successor with its posterior distribution conditional on the forecast. We use the same data, discretization, battery dynamics, and evaluation interval as Lu et al. (2025). The MDP has 2,100 states, 9 actions, and $\gamma = 0.95$. At date t, the current price and mismatch are observed exactly, with predictions for $t+1, \ldots, t+\ell$, where $\ell \in \{0,1,2,3\}$. Forecasts are perturbed as $\widehat{x}_u = x_u(1+\sigma\xi_u)$, where x_u denotes either the electricity price or the production mismatch at date u, and $\xi_u \sim \mathcal{N}(0,1)$

with $\sigma \in \{0, .05, .10, .20, .30, .40, .50, .60, .80, 1\}$. Following their convention, noise is reported as $100\sigma\%$. RPTAS converts these forecasts into posterior kernels and uses $N = 8$ dictionary samples.

We compare RPTAS with three controllers. BOLA is the zero-terminal, point-forecast controller used in the wind-farm experiments of Lu et al. (2025). BOLA-BAYES follows the method described in their paper: it uses forecast posteriors and a Bayesian continuation value computed offline. Both execute a block of $\ell+1$ actions before replanning. We also include MPC, which uses the same finite-horizon objective and zero terminal value as BOLA, but replans after every transition. Comparing MPC with BOLA isolates the effect of replanning. Comparing it with RPTAS shows why replanning alone need not suffice: RPTAS also accounts for the value beyond the forecast window through its dictionary value function. All methods face the same realised trajectory, initial battery state, constraints, and forecast-noise draws.

We report realised cost reduction relative to the no-prediction controller used by Lu et al. (2025). The no-prediction controller plans as if the current price and production mismatch remained fixed; it defines the grey reference line at zero. The red dashed line instead shows the optimal zero-look-ahead policy for our estimated MDP ($\ell = 0$). Higher cost reduction is better. Curves average ten common forecast-noise seeds, with two-standard-error bands, see Section B for more details on the implementation.

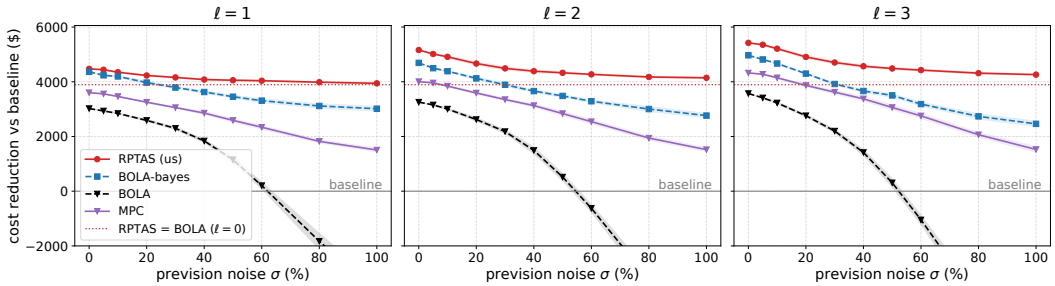


Figure 1: Realised cost reduction relative to the no-prediction baseline versus forecast noise σ, for $\ell \in \{1, 2, 3\}$. The grey and red lines denote the no-prediction and classical MDP ($\ell = 0$) benchmarks, respectively. Curves show means over 10 seeds, with ± 2 standard-error bands.

7 CONCLUSION

We investigated whether multi-step transition look-ahead can be used efficiently. Although exact planning is NP-hard for every fixed rational discount factor and every fixed depth $\ell \geq 2$, a uniformly near-optimal policy can be computed in polynomial time for fixed ℓ. Our algorithm samples a polynomial number N of transition tables and runs value iteration on $\mathcal{S} \times [N]^{\ell}$, rather than on the exponentially large space $\mathcal{S} \times \Omega^{\ell}$. A backward recursion then uses the resulting values to act on *any* observed look-ahead window, yielding a uniformly near-optimal policy with polynomial preprocessing and per-decision cost for fixed ℓ. With unknown dynamics, the same representation yields a learning algorithm whose leading regret term is comparable to that of classical discounted RL. We empirically validate the soundness of our results on the wind-farm storage-control benchmark of Lu et al. (2025), showing that our approach, optimally accounting for ℓ-step look-ahead information, offers substantially better performance than existing algorithms. A natural direction is to move beyond the local corruption model and study forecasts whose errors are biased or correlated across state–action pairs and over time, as well as forecasts that reveal only partial information about future transitions. Also, in practice, such predictions may be continuous, high-dimensional, and generated by a learned world model. Extending the dictionary approach to this setting, while relating planning performance and regret to forecast quality, is an important step toward practical forecast-aware reinforcement learning.

AI USE STATEMENT

In this work, we used generative AI tools to implement code for our experiments. Additionally, we used generative AI tools to help draft parts of the manuscript. We have reviewed all AI-assisted

work: LLM-generated code was verified and tested for correctness by the authors before being used in our experiments, and LLM-assisted text was checked and revised by the authors for accuracy and originality. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rohit Agrawal. Finite-sample concentration of the multinomial in relative entropy. *IEEE Transactions on Information Theory*, 66(10):6297–6302, October 2020. ISSN 1557-9654. doi: 10.1109/tit.2020.2996134. URL http://dx.doi.org/10.1109/TIT.2020.2996134.

Mohammad Gheshlaghi Azar, Remi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model, 2012. URL https://arxiv.org/abs/1206.6461.

Nikhil Balaji, Stefan Kiefer, Petr Novotnỳ, Guillermo A Pérez, and Mahsa Shirmohammadi. On the complexity of value iteration. *arXiv preprint arXiv:1807.04920*, 2018.

Yichen Chen and Mengdi Wang. Lower bound on the computational complexity of discounted markov decision problems, 2017. URL https://arxiv.org/abs/1705.07312.

Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine Learning*, 49(2–3):193–208, 2002. doi: 10.1023/A:1017932429737.

Anton J. Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002. doi: 10.1137/S1052623499363220.

Tongxin Li, Yiheng Lin, Shaolei Ren, and Adam Wierman. Beyond black-box advice: Learning-augmented algorithms for MDPs with q-value predictions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=RACcp8Zbr9.

Tongxin Li, Hao Liu, and Yisong Yue. Disentangling linear quadratic control with untrusted ml predictions. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 86860–86898. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9dff3b83d463fab213941bfee23341ba-Paper-Conference.pdf.

Michael L. Littman, Thomas L. Dean, and Leslie Pack Kaelbling. On the complexity of solving markov decision problems, 2013. URL https://arxiv.org/abs/1302.4971.

Shuang Liu and Hao Su. Regret bounds for discounted mdps, 2021. URL https://arxiv.org/abs/2002.05138.

Chenbei Lu, Zaiwei Chen, Tongxin Li, Chenye Wu, and Adam Wierman. Reinforcement learning with imperfect transition predictions: A bellman-jensen approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=DYuPwwDy9n.

Lixing Lyu, Jiashuo Jiang, and Wang Chi Cheung. Efficiently solving discounted MDPs via predictions with unknown prediction errors. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=0nrxgFZEEq.

Davide Maran, Davide Salaorni, and Marcello Restelli. Learning in markov decision processes with exogenous dynamics, 2026. URL https://arxiv.org/abs/2603.02862.

Aranyak Mehta, Uri Nadav, Alexandros Psomas, and Aviad Rubinstein. Hitting the high notes: Subset selection for maximizing expected order statistics. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15800–15810. Curran Associates, Inc.,

2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b6417f112bd27848533e54885b66c288-Paper.pdf.

Nadav Merlis. Reinforcement learning with lookahead information. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=wlqfOvlTQz.

Nadav Merlis, Dorian Baudry, and Vianney Perchet. The value of reward lookahead in reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2403.11637.

Martin Mundhenk, Judy Goldsmith, Christopher Lusena, and Eric Allender. Complexity of finite-horizon markov decision process problems. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, pp. 494–499, 2000.

C. H. Papadimitriou. The complexity of markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.

Corentin Pla, Hugo Richard, Marc Abeille, Nadav Merlis, and Vianney Perchet. On the hardness of reinforcement learning with transition lookahead. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026a. URL https://openreview.net/forum?id=clyOoEL3pS.

Corentin Pla, Hugo Richard, Marc Abeille, and Vianney Perchet. Minimax pac bounds for learning in exogenous contextual mdps, 2026b. URL https://arxiv.org/abs/2606.25170.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018.

Thomas J. Walsh, Ali Nouri, Lihong Li, and Michael L. Littman. Learning and planning in environments with delayed feedback. *Autonomous Agents and Multi-Agent Systems*, 18(1):83–105, 2009. doi: 10.1007/s10458-008-9056-7.

A Appendix

A.1 Proof of Theorem 1

We use the notation of Section 2: $\Omega = \mathcal{S}^{\mathcal{S}\times\mathcal{A}}$ is the space of one-step transition tables, L is the product law equation 1 induced by the kernel P, and $V_\ell^\star$ denotes the optimal value function on the augmented state space $\mathcal{S}\times\Omega^\ell$. For a policy π and an initial state $s_0 \in \mathcal{S}$, define the values averaged over the initial look-ahead window,

$$v^\pi_{\ell,\gamma}(s_0) := \mathbb{E}_{C\sim L^{\otimes\ell}}\big[V^\pi_\ell(s_0, C)\big], \qquad v^\star_{\ell,\gamma}(s_0) := \mathbb{E}_{C\sim L^{\otimes\ell}}\big[V^\star_\ell(s_0, C)\big], \tag{13}$$

For a window $c = (\theta_0, \dots, \theta_{\ell-1}) \in \Omega^\ell$, a state $s \in \mathcal{S}$, and an action sequence $\sigma = (a_0, \dots, a_{k-1})$ with $k \le \ell$, we write $\phi_c(s, \sigma) \in \mathcal{S}$ for the state reached after playing σ from s under the window c, that is, the state s_k obtained from the recursion

$$s_0 = s, \qquad s_{j+1} = \theta_j(s_j, a_j), \qquad j = 0, \dots, k-1.$$

Definition 2 (Discounted Value Decision Problem (ℓ-DVDP)). *Given a finite MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$, a rational discount factor $\gamma \in (0,1)$, an initial state $s_0 \in \mathcal{S}$, and a rational threshold θ, decide whether there exists a policy π with ℓ-step transition look-ahead such that $v^\pi_{\ell,\gamma}(s_0) \ge \theta$, equivalently whether $v^\star_{\ell,\gamma}(s_0) \ge \theta$.*

Fix an integer $\ell \ge 2$ and a rational discount factor $\gamma \in (0,1)$. We build upon Pla et al. (2026a) and exhibit a polynomial-time reduction from Independent Set on regular graphs to ℓ-DVDP that is valid for this fixed γ; NP-hardness is therefore preserved even for γ far from 1. We use the following consequence of the reduction of Mehta et al. (2020); it is the same gap gadget as in the hardness proof of Pla et al. (2026a). The graph parameters are denoted n_G and m_G.

Lemma 1 (Expected-maximum gap)**.** *There is a polynomial-time reduction that maps an instance (G, k) of* INDEPENDENT SET *on a regular graph $G = (V, E)$, with $n_G = |V|$ and $m_G = |E|$, to mutually independent, nonnegative, finite-support random variables $(X_v)_{v \in V}$ with rational values and probabilities of polynomial encoding length. All variables have the same expectation μ, and the following hold:*

1. *If G contains an independent set $S^\star$ of size k, then*
$$\mathbb{E}\left[\max_{v \in S^\star} X_v\right] \geq k\mu - \frac{2}{m_G}.$$
2. *If G contains no independent set of size k, then every $S \subseteq V$ with $|S| = k$ satisfies*
$$\mathbb{E}\left[\max_{v \in S} X_v\right] \leq k\mu - 1.$$

We may assume without loss of generality that $m_G \geq 3$: instances with at most two edges can be decided in polynomial time. Set
$$\Delta := 1 - \frac{2}{m_G} > 0.$$
Define
$$B := k\mu - 1, \qquad M := 1 + \max\{0, -B\}, \qquad U := M + B.$$
Then $U \geq 1$. Shift every random variable by the same deterministic amount:
$$Y_v := X_v + M.$$
For every nonempty $S \subseteq V$,
$$\mathbb{E}\left[\max_{v \in S} Y_v\right] = M + \mathbb{E}\left[\max_{v \in S} X_v\right].$$
Consequently, Lemma 1 becomes

$$\text{NO case:} \quad \mathbb{E}\left[\max_{v \in S} Y_v\right] \leq U \qquad \text{for every } S \subseteq V,\ |S| = k, \tag{14}$$

$$\text{YES case:} \quad \mathbb{E}\left[\max_{v \in S^\star} Y_v\right] \geq U + \Delta \qquad \text{for some independent set } S^\star,\ |S^\star| = k. \tag{15}$$

The shift is used only to ensure that all rewards introduced below are nonnegative.

MDP construction. For each $v \in V$, write the finite support of Y_v as
$$\operatorname{supp}(Y_v) = \{y_{v,1}, \ldots, y_{v,N_v}\}, \qquad \mathbb{P}(Y_v = y_{v,h}) = p_{v,h}.$$
The state space of the MDP $\mathcal{M}_G$ is
$$\begin{aligned} \mathcal{S} = {} & \{s_0, s_1, s_T\} \cup \{d_1, \ldots, d_{\ell-2}\} \cup \{s_v : v \in V\} \\ & \cup \{x_{v,h} : v \in V,\ h \in [N_v]\}. \end{aligned} \tag{16}$$
The delay states $d_1, \ldots, d_{\ell-2}$ are absent when $\ell = 2$. The action set is
$$\mathcal{A} = \{\mathsf{wait}, \mathsf{go}, \mathsf{advance}, \mathsf{claim}, \mathsf{collect}\} \cup \{\mathsf{pick}_1, \ldots, \mathsf{pick}_k\}. \tag{17}$$
Set
$$T := \gamma^{\ell+1} U, \qquad b := (1 - \gamma)T. \tag{18}$$
The only nonzero rewards are
$$r(s_0, \mathsf{wait}) = b, \qquad r(x_{v,h}, \mathsf{collect}) = y_{v,h}.$$
All rewards are nonnegative and bounded by $\max_{v,h} y_{v,h}$.

The transition kernel is as follows.

1. At the root s_0,
$$P(s_0 \mid s_0, \mathsf{wait}) = 1.$$
If $\ell = 2$, action go moves deterministically to s_1; if $\ell \geq 3$, it moves deterministically to d_1.
2. At a delay state d_i, action advance moves deterministically to d_{i+1} when $i < \ell - 2$, and from $d_{\ell-2}$ to s_1.
3. At the selector state s_1, each action pick_j, $j \in [k]$, has the uniform transition law
$$P(s_v \mid s_1, \mathsf{pick}_j) = \frac{1}{n_G}, \qquad v \in V. \tag{19}$$
4. At a vertex state s_v, action claim samples the corresponding payoff state:
$$P(x_{v,h} \mid s_v, \mathsf{claim}) = p_{v,h}, \qquad h \in [N_v].$$
5. At a payoff state $x_{v,h}$, action collect moves deterministically to s_T.
6. The terminal state s_T is absorbing. Every action not explicitly specified above moves deterministically to s_T and gives reward zero.

Candidate tuples and the value of committing. We call *root augmented state* any augmented state of the form (s_0, c) with $c = (\theta_0, \ldots, \theta_{\ell-1}) \in \Omega^\ell$. For $j \in [k]$, define the length-ℓ action sequence
$$\sigma_j := (\mathsf{go}, \underbrace{\mathsf{advance}, \ldots, \mathsf{advance}}_{\ell-2 \text{ times}}, \mathsf{pick}_j), \tag{20}$$
with the obvious interpretation $(\mathsf{go}, \mathsf{pick}_j)$ when $\ell = 2$. The first $\ell - 1$ actions of σ_j deterministically drive the system from s_0 to s_1, so there is a unique $q_j(c) \in V$ such that
$$\phi_c(s_0, \sigma_j) = s_{q_j(c)}, \qquad \text{namely } s_{q_j(c)} = \theta_{\ell-1}(s_1, \mathsf{pick}_j).$$
Write
$$q(c) := (q_1(c), \ldots, q_k(c)) \in V^k, \qquad S(c) := \{q_1(c), \ldots, q_k(c)\}. \tag{21}$$
When $C \sim L^{\otimes \ell}$, the coordinates of $q(C)$ are independent and uniform on V, because they arise from the distinct state–action pairs (s_1, pick_j) of a single product-distributed table, via equation 19.

Lemma 2 (Commit value and root recursion)**.** *For every root augmented state (s_0, c), the value obtained by choosing* go *at (s_0, c) and then acting optimally is*
$$C(c) = \gamma^{\ell+1} \mathbb{E}\left[\max_{v \in S(c)} Y_v\right]. \tag{22}$$
Moreover,
$$V_\ell^\star(s_0, c) = \max\left\{C(c),\ b + \gamma\, \mathbb{E}[V_\ell^\star(s_0, C') \mid c, \mathsf{wait}]\right\}, \tag{23}$$
where $C' = (\theta_1, \ldots, \theta_{\ell-1}, \Theta)$ with $\Theta \sim L$ independent of the past. Finally, after playing wait*, the next candidate tuple $q(C')$ is an independent fresh draw from the uniform distribution on V^k.*

Proof. After go, the process traverses the deterministic delay chain and cannot return to s_0. There are no rewards on this part of the trajectory. The system reaches s_1 exactly $\ell - 1$ steps after leaving s_0. At that time, the current ℓ-step look-ahead window contains, for every $j \in [k]$, both
$$s_{q_j(c)} \quad \text{and} \quad x_{q_j(c),h_j}$$
along the continuation $(\mathsf{pick}_j, \mathsf{claim})$, since $\ell \geq 2$. Hence the realization of $Y_{q_j(c)}$ is known before the action pick_j is selected. Note also that, by prefix consistency of the look-ahead windows, the vertex reached under pick_j at that time is exactly $q_j(c)$: the transition (s_1, pick_j) is governed by the last table of the root window, which is also the table governing the system when it actually reaches s_1. For distinct vertices, these payoff samples are independent; if the same vertex appears several times in $q(c)$, prefix consistency makes the corresponding branches share the same sample from the pair (s_v, claim). Therefore the selector obtains the largest realized value among the distinct variables indexed by $S(c)$.

Moreover, the payoff realizations are not observable at the root: the payoff states $x_{v,h}$ lie at depth $\ell+1$ from s_0 under $(\sigma_j, \mathsf{claim})$, one level beyond the depth-ℓ tree revealed by c. Hence, conditionally on the root window, the variables $(Y_v)_{v\in S(c)}$ have their unconditional product law.

Starting from s_1, the reward is collected two steps later: one transition under pick_j, one under claim, and then the immediate reward of $\mathsf{collect}$ at the payoff state. The conditional value at s_1 is thus

$$\gamma^2 \max_{v\in S(c)} Y_v.$$

The selector is reached $\ell - 1$ steps after leaving s_0, so the value at the root is

$$\gamma^{\ell-1}\,\mathbb{E}\left[\gamma^2 \max_{v\in S(c)} Y_v\right] = \gamma^{\ell+1}\,\mathbb{E}\left[\max_{v\in S(c)} Y_v\right],$$

which proves equation 22.

At s_0, the only potentially optimal actions are go and wait, since every other action leads to the absorbing state s_T with zero reward. The former has value $C(c)$. The latter gives immediate reward b, returns to s_0, and then has continuation value $\gamma\,\mathbb{E}[V_\ell^\star(s_0, C') \mid c, \mathsf{wait}]$. This proves equation 23.

Finally, after wait, the candidate transitions lie one level beyond the old look-ahead tree: $q(C')$ is determined by the entries (s_1, pick_j) of the newly revealed table Θ, which is independent of all previously observed tables. The augmented transition kernel therefore samples them independently from equation 19, independently of the previous tuple. Hence $q(C')$ is a fresh uniform draw from V^k. In particular, the current window never reveals the next candidate tuple before wait is played. □

Proof of Theorem 1. We show that the map $(G, k) \mapsto (\mathcal{M}_G, s_0, \gamma, \theta)$, with θ defined in equation 31 below, is a valid many-one reduction from INDEPENDENT SET to ℓ-DVDP.

Soundness. Assume that G is a NO instance of INDEPENDENT SET. Fix a root window $c \in \Omega^\ell$. The set $S(c)$ may contain fewer than k vertices because the tuple can have repetitions. Extend it to any set $\widetilde{S} \supseteq S(c)$ of cardinality k. Since the expected maximum is monotone under set inclusion (the Y_v are nonnegative), equation 14 gives

$$\mathbb{E}\left[\max_{v\in S(c)} Y_v\right] \le \mathbb{E}\left[\max_{v\in \widetilde{S}} Y_v\right] \le U. \tag{24}$$

By equation 22 and equation 18,

$$C(c) \le \gamma^{\ell+1} U = T. \tag{25}$$

Let

$$M := \max\{V_\ell^\star(s_0, c) : c \in \Omega^\ell\}.$$

The set Ω^ℓ is finite, so the maximum exists. From Lemma 2 and equation 25,

$$M \le \max\{T,\ b + \gamma M\}.$$

If $M > T$, then necessarily

$$M \le b + \gamma M = (1-\gamma)T + \gamma M,$$

which implies $(1-\gamma)M \le (1-\gamma)T$, contradicting $M > T$. Therefore

$$M \le T. \tag{26}$$

In particular, after averaging over the initial root window,

$$v_{\ell,\gamma}^\star(s_0) \le T. \tag{27}$$

Completeness. Assume that G is a YES instance, and let

$$S^\star = \{v_1, \dots, v_k\}$$

be an independent set satisfying equation 15. Consider the following stationary policy on the augmented state space:

- at a root augmented state (s_0, c), choose go if
$$q(c) = (v_1, \dots, v_k),$$
and choose wait otherwise;
- traverse the deterministic delay chain using advance;
- at s_1, choose an index whose revealed payoff is maximal;
- use claim and then collect.

This policy is well defined: the tuple $q(c)$ is observable at the root, and the realized payoffs are observable at s_1, by Lemma 2.

At every visit to s_0, the target tuple appears with probability

$$\rho := n_G^{-k}. \tag{28}$$

By the last part of Lemma 2, successive tuples are independent. Let $\tau \in \{0, 1, 2, \dots\}$ be the number of waiting actions before the target tuple first appears. Then

$$\mathbb{P}(\tau = t) = (1-\rho)^t \rho,$$

and therefore

$$\alpha := \mathbb{E}[\gamma^\tau] = \sum_{t=0}^{\infty} \rho(1-\rho)^t \gamma^t = \frac{\rho}{1-\gamma(1-\rho)}. \tag{29}$$

The discounted return of this policy is

$$\sum_{t=0}^{\tau-1} \gamma^t b + \gamma^{\tau+\ell+1} \max_{v \in S^\star} Y_v.$$

The payoff samples generated after commitment are independent of the waiting time (they are drawn from table entries never used to define the tuples) and have the laws $(Y_v)_{v \in S^\star}$, with all vertices of the target tuple distinct. Taking expectations and using $b/(1-\gamma) = T$ gives

$$\begin{aligned} v^\pi_{\ell,\gamma}(s_0) &= T\big(1 - \mathbb{E}[\gamma^\tau]\big) + \mathbb{E}[\gamma^\tau]\, \gamma^{\ell+1}\, \mathbb{E}\left[\max_{v \in S^\star} Y_v\right] \\ &= T + \alpha \left(\gamma^{\ell+1}\, \mathbb{E}\left[\max_{v \in S^\star} Y_v\right] - T \right) \\ &\geq T + \alpha \gamma^{\ell+1} \Delta, \end{aligned} \tag{30}$$

where the last inequality follows from equation 15 and $T = \gamma^{\ell+1} U$.

Decision threshold and polynomial encoding. Define

$$\theta := T + \frac{1}{2} \alpha \gamma^{\ell+1} \Delta. \tag{31}$$

Because $\gamma > 0$, $\rho > 0$, and $\Delta > 0$, the added term is strictly positive. Equations equation 27 and equation 30 imply

$$\begin{aligned} G \text{ is a NO instance} \quad &\Longrightarrow \quad v^\star_{\ell,\gamma}(s_0) \leq T < \theta, \\ G \text{ is a YES instance} \quad &\Longrightarrow \quad v^\star_{\ell,\gamma}(s_0) \geq T + \alpha \gamma^{\ell+1} \Delta > \theta. \end{aligned}$$

Thus $(G, k) \mapsto (\mathcal{M}_G, s_0, \gamma, \theta)$ is a valid many-one reduction to ℓ-DVDP.

It remains to verify polynomial size. The stochastic gadget in Lemma 1 has polynomially many support points and polynomial-bit rational values and probabilities. The MDP has

$$O\left(n_G + \sum_{v \in V} N_v + \ell\right)$$

states and $k + 5$ actions. Since ℓ is fixed, this is polynomial. The shift L, the quantities U, T, b, Δ, and the threshold θ are obtained by a constant number of rational arithmetic operations. Moreover, $\rho = n_G^{-k}$ has encoding length $O(k \log n_G)$, and equation 29 therefore has polynomial encoding length. For fixed rational γ, its encoding length is constant; if γ is supplied in binary, all constructed numbers still have encoding length polynomial in the combined input size. This completes the reduction and the proof. □

A.2 Proof of Theorem 2

Notation and conventions. We use the notation of Section 2: $\Omega = \mathcal{S}^{\mathcal{S}\times\mathcal{A}}$, L is the law equation 1 on Ω, $\mathcal{T}_\ell$ is the Bellman optimality operator of the augmented MDP, and $V_\ell^\star$, $Q_\ell^\star$ are its optimal value and action-value functions on $\mathcal{S} \times \Omega^\ell$. We write $V_{\max} := (1-\gamma)^{-1}$, so that every value function below takes values in $[0, V_{\max}]$, and $\theta_{k:k'} := (\theta_k, \dots, \theta_{k'})$ for windows. Throughout, s denotes the root state of a query, u an intermediate state, a the root action, a' an intermediate action, k a value-iteration index, $m \in \{1, \dots, \ell\}$ a level of the backward recursion, and $i, j \in [N] := \{1, \dots, N\}$ dictionary indices. The statement of Theorem 2 assumes $\ell \geq 2$ because exact planning is already polynomial for $\ell = 1$; the proof below is valid for every $\ell \geq 1$.

The proof has three steps: (i) an *ideal empirical model*, defined on the full space $\mathcal{S} \times \Omega^\ell$, is shown to be uniformly close to the true model; (ii) its optimal value is shown to coincide, on windows made of dictionary tables, with the fixed point of the finite value iteration of Algorithm 1; (iii) the backward recursion of Algorithm 2 is shown to reproduce the empirical action scores exactly on *any* window.

(i) Sample once and for all

$$\Theta^1, \dots, \Theta^N \overset{\text{i.i.d.}}{\sim} L, \tag{32}$$

and let

$$\widehat{L}_N := \frac{1}{N} \sum_{j=1}^{N} \delta_{\Theta^j} \tag{33}$$

denote the empirical distribution of the dictionary, as in Section 4. The ideal empirical Bellman operator acts on $V : \mathcal{S} \times \Omega^\ell \to \mathbb{R}$ by

$$(\widehat{\mathcal{T}}_\ell V)(s, \theta_{0:\ell-1}) := \max_{a \in \mathcal{A}} \Big\{ r(s,a) + \frac{\gamma}{N} \sum_{j=1}^{N} V\big(\theta_0(s,a), \theta_{1:\ell-1}, \Theta^j\big) \Big\}. \tag{34}$$

It is a γ-contraction for the sup norm on $\mathcal{S} \times \Omega^\ell$; let $\widehat{V}^\star$ be its unique fixed point. Recall the true action-value function and define its empirical counterpart:

$$Q_\ell^\star(s, \theta_{0:\ell-1}, a) = r(s,a) + \gamma\, \mathbb{E}_{\Theta \sim L}\big[V_\ell^\star\big(\theta_0(s,a), \theta_{1:\ell-1}, \Theta\big)\big], \tag{35}$$

$$\widehat{Q}^\star(s, \theta_{0:\ell-1}, a) := r(s,a) + \frac{\gamma}{N} \sum_{j=1}^{N} \widehat{V}^\star\big(\theta_0(s,a), \theta_{1:\ell-1}, \Theta^j\big). \tag{36}$$

To compare the two models, consider the function class

$$\mathcal{F} := \big\{ \Theta \mapsto V_\ell^\star(u, \theta_{1:\ell-1}, \Theta) : u \in \mathcal{S},\ \theta_{1:\ell-1} \in \Omega^{\ell-1} \big\}, \tag{37}$$

whose cardinality satisfies $|\mathcal{F}| \leq |\mathcal{S}||\Omega|^{\ell-1} = |\mathcal{S}|^{1+(\ell-1)|\mathcal{S}||\mathcal{A}|}$. Crucially, $\log|\mathcal{F}| = \big(1 + (\ell-1)|\mathcal{S}||\mathcal{A}|\big) \log|\mathcal{S}|$ is polynomial in $|\mathcal{S}|$ and $|\mathcal{A}|$ for fixed ℓ.

Lemma 3 (Uniform concentration)**.** *For every $\eta > 0$,*

$$\mathbb{P}\left(\sup_{u, \theta_{1:\ell-1}} \left| \frac{1}{N} \sum_{j=1}^{N} V_\ell^\star(u, \theta_{1:\ell-1}, \Theta^j) - \mathbb{E}_{\Theta \sim L}\big[V_\ell^\star(u, \theta_{1:\ell-1}, \Theta)\big] \right| > \eta \right) \leq 2|\mathcal{S}||\Omega|^{\ell-1} \exp\left(-\frac{2N\eta^2}{V_{\max}^2} \right). \tag{38}$$

Proof. $V_\ell^\star$ is deterministic and does not depend on the sampled dictionary, so each function in $\mathcal{F}$ is fixed and takes values in $[0, V_{\max}]$. Apply Hoeffding's inequality to each of them and take a union bound over $\mathcal{F}$. □

Denote by $\mathcal{E}_\eta$ the complement of the event in equation 38, i.e. the event on which the supremum is at most η.

Lemma 4 (Fixed-point stability)**.** *On $\mathcal{E}_\eta$,*

$$\|\widehat{V}^\star - V_\ell^\star\|_\infty \le \frac{\gamma\eta}{1-\gamma}. \tag{39}$$

Proof. Fix $(s, \theta_{0:\ell-1})$. For every $a \in \mathcal{A}$, the bracketed terms in $\widehat{\mathcal{T}}_\ell V_\ell^\star$ and $\mathcal{T}_\ell V_\ell^\star$ share the reward $r(s,a)$ and differ by γ times an empirical-versus-true average of $V_\ell^\star(\theta_0(s,a), \theta_{1:\ell-1}, \cdot) \in \mathcal{F}$, which is at most $\gamma\eta$ in absolute value on $\mathcal{E}_\eta$. Since the maximum over a is 1-Lipschitz, $\|\widehat{\mathcal{T}}_\ell V_\ell^\star - \mathcal{T}_\ell V_\ell^\star\|_\infty \le \gamma\eta$. As $V_\ell^\star = \mathcal{T}_\ell V_\ell^\star$, $\widehat{V}^\star = \widehat{\mathcal{T}}_\ell \widehat{V}^\star$, and $\widehat{\mathcal{T}}_\ell$ is a γ-contraction,

$$\|\widehat{V}^\star - V_\ell^\star\|_\infty \le \|\widehat{\mathcal{T}}_\ell \widehat{V}^\star - \widehat{\mathcal{T}}_\ell V_\ell^\star\|_\infty + \|\widehat{\mathcal{T}}_\ell V_\ell^\star - \mathcal{T}_\ell V_\ell^\star\|_\infty \le \gamma\|\widehat{V}^\star - V_\ell^\star\|_\infty + \gamma\eta,$$

and rearranging proves the claim. □

Lemma 5 (Action-value perturbation)**.** *On $\mathcal{E}_\eta$,*

$$\sup_{s,\theta_{0:\ell-1},a} \left|\widehat{Q}^\star(s,\theta_{0:\ell-1},a) - Q_\ell^\star(s,\theta_{0:\ell-1},a)\right| \le \frac{\gamma\eta}{1-\gamma}. \tag{40}$$

Proof. Adding and subtracting the empirical average of $V_\ell^\star$,

$$\begin{aligned}
&\left|\widehat{Q}^\star(s,\theta_{0:\ell-1},a) - Q_\ell^\star(s,\theta_{0:\ell-1},a)\right| \\
&\le \frac{\gamma}{N}\sum_{j=1}^{N}\left|\widehat{V}^\star\big(\theta_0(s,a),\theta_{1:\ell-1},\Theta^j\big) - V_\ell^\star\big(\theta_0(s,a),\theta_{1:\ell-1},\Theta^j\big)\right| + \gamma\eta \\
&\le \gamma\|\widehat{V}^\star - V_\ell^\star\|_\infty + \gamma\eta \le \frac{\gamma\eta}{1-\gamma},
\end{aligned}$$

where the last step uses Lemma 4 and $\gamma \cdot \frac{\gamma\eta}{1-\gamma} + \gamma\eta = \frac{\gamma\eta}{1-\gamma}$. □

(ii) The function $\widehat{V}^\star$ is close to $V_\ell^\star$ but is still defined on the exponentially large space $\mathcal{S} \times \Omega^\ell$. We now show that windows composed only of dictionary tables form a finite state space preserved by $\widehat{\mathcal{T}}_\ell$, on which $\widehat{V}^\star$ can be computed by ordinary value iteration. The results of this step and of Step 3 hold for *any* fixed dictionary $(\Theta^1, \dots, \Theta^N) \in \Omega^N$, whatever its law.

Recall the proxy state space $\mathcal{D}_N := \mathcal{S} \times [N]^\ell$ of Section 4 and define the embedding

$$\iota : \mathcal{D}_N \to \mathcal{S} \times \Omega^\ell, \qquad \iota(s, i_0, \dots, i_{\ell-1}) := (s, \Theta^{i_0}, \dots, \Theta^{i_{\ell-1}}), \tag{41}$$

which reads an index tuple as the window of dictionary tables it encodes, the index i_m specifying the table at position m of the window. For $V : \mathcal{S} \times \Omega^\ell \to \mathbb{R}$, let $R_N V := V \circ \iota : \mathcal{D}_N \to \mathbb{R}$ denote its restriction to dictionary windows. The proxy Bellman operator acts on arrays $v : \mathcal{D}_N \to \mathbb{R}$ by

$$(\widehat{\mathcal{T}}_N v)(s, i_0, \dots, i_{\ell-1}) := \max_{a\in\mathcal{A}}\Big\{ r(s,a) + \frac{\gamma}{N}\sum_{j=1}^{N} v\big(\Theta^{i_0}(s,a), i_1, \dots, i_{\ell-1}, j\big)\Big\}; \tag{42}$$

this is exactly the update equation 6 of Algorithm 1. It is a γ-contraction on $\mathbb{R}^{\mathcal{D}_N}$; let $V_N^\star$ denote its unique fixed point.

Lemma 6 (Exact restriction to the dictionary)**.** *$V_N^\star = R_N \widehat{V}^\star$, that is, for every $(s, i_0, \dots, i_{\ell-1}) \in \mathcal{D}_N$,*

$$V_N^\star(s, i_0, \dots, i_{\ell-1}) = \widehat{V}^\star(s, \Theta^{i_0}, \dots, \Theta^{i_{\ell-1}}). \tag{43}$$

Proof. We first show that $R_N\widehat{\mathcal{T}}_\ell V = \widehat{\mathcal{T}}_N R_N V$ for every $V : \mathcal{S} \times \Omega^\ell \to \mathbb{R}$. Fix $(s, i_0, \dots, i_{\ell-1}) \in \mathcal{D}_N$. By equation 34 and equation 41,

$$\begin{aligned}(R_N\widehat{\mathcal{T}}_\ell V)(s, i_0, \dots, i_{\ell-1}) &= (\widehat{\mathcal{T}}_\ell V)(s, \Theta^{i_0}, \dots, \Theta^{i_{\ell-1}}) \\ &= \max_{a\in\mathcal{A}}\Big\{r(s,a) + \frac{\gamma}{N}\sum_{j=1}^{N} V\big(\Theta^{i_0}(s,a), \Theta^{i_1}, \dots, \Theta^{i_{\ell-1}}, \Theta^{j}\big)\Big\} \\ &= \max_{a\in\mathcal{A}}\Big\{r(s,a) + \frac{\gamma}{N}\sum_{j=1}^{N} (R_N V)\big(\Theta^{i_0}(s,a), i_1, \dots, i_{\ell-1}, j\big)\Big\} \\ &= (\widehat{\mathcal{T}}_N R_N V)(s, i_0, \dots, i_{\ell-1}),\end{aligned}$$

where the third equality holds because the window $(\Theta^{i_1}, \dots, \Theta^{i_{\ell-1}}, \Theta^{j})$ consists of dictionary tables. Applying this identity to $V = \widehat{V}^\star$ and using $\widehat{\mathcal{T}}_\ell\widehat{V}^\star = \widehat{V}^\star$ gives $\widehat{\mathcal{T}}_N(R_N\widehat{V}^\star) = R_N(\widehat{\mathcal{T}}_\ell\widehat{V}^\star) = R_N\widehat{V}^\star$. Thus $R_N\widehat{V}^\star$ is a fixed point of the γ-contraction $\widehat{\mathcal{T}}_N$, whose fixed point is unique; hence $R_N\widehat{V}^\star = V_N^\star$. □

Algorithm 1 computes $V_N^\star$ approximately by value iteration,

$$V^0 := 0, \qquad V^{k+1} := \widehat{\mathcal{T}}_N V^k, \qquad k = 0, \dots, K-1, \tag{44}$$

and by contraction $\|V^K - V_N^\star\|_\infty \le \gamma^K V_{\max}$.

Remark 1 (Cost and memory of one iteration)**.** *For an array $v : \mathcal{D}_N \to \mathbb{R}$ and every $(u, i_1, \dots, i_{\ell-1}) \in \mathcal{S} \times [N]^{\ell-1}$, first compute*

$$\mathrm{Mean}_v(u, i_1, \dots, i_{\ell-1}) := \frac{1}{N}\sum_{j=1}^{N} v(u, i_1, \dots, i_{\ell-1}, j), \tag{45}$$

which costs $O(|\mathcal{S}|N^\ell)$ in total; then each of the $|\mathcal{S}|N^\ell$ maxima in equation 42 costs $O(|\mathcal{A}|)$ look-ups into Mean_v. One iteration of equation 42 therefore costs

$$O\big(|\mathcal{S}|(|\mathcal{A}|+1)N^\ell\big), \tag{46}$$

and the arrays v and Mean_v occupy $O(|\mathcal{S}|N^\ell)$ memory.

(iii) $V_N^\star$ is only defined on dictionary windows, whereas the window observed at deployment is an arbitrary $c = (\theta_0, \dots, \theta_{\ell-1}) \in \Omega^\ell$. The key observation is that after ℓ transitions every table of c has been shifted out and replaced by a fresh table, which the empirical model draws from the dictionary. Algorithm 2 therefore solves an ℓ-step planning problem with terminal values on $\mathcal{D}_N$, reintroducing the observed tables $\theta_{\ell-1}, \dots, \theta_0$ by backward recursion. We define this recursion for an arbitrary terminal array so that we can analyse both the exact array $V_N^\star$ and the computed array V^K.

Fix a window $c = (\theta_0, \dots, \theta_{\ell-1}) \in \Omega^\ell$ and an array $v : \mathcal{D}_N \to \mathbb{R}$. Define $U^v_{m,c} : \mathcal{S} \times [N]^m \to \mathbb{R}$ for $m = \ell, \ell-1, \dots, 1$ by the terminal condition

$$U^v_{\ell,c}(u, i_0, \dots, i_{\ell-1}) := v(u, i_0, \dots, i_{\ell-1}), \tag{47}$$

and, for $m = \ell-1, \dots, 1$,

$$U^v_{m,c}(u, i_0, \dots, i_{m-1}) := \max_{a'\in\mathcal{A}}\Big\{r(u,a') + \frac{\gamma}{N}\sum_{i_m=1}^{N} U^v_{m+1,c}\big(\theta_m(u,a'), i_0, \dots, i_m\big)\Big\}. \tag{48}$$

Here $U^v_{m,c}(u, i_0, \dots, i_{m-1})$ is the continuation value at step m of the query, when the remaining observed tables are $\theta_{m:\ell-1}$ and the m tables revealed since the query are represented by the dictionary indices $i_0, \dots, i_{m-1}$, in order of arrival. Finally, the action scores at the root are

$$\widetilde{Q}_v(s, c, a) := r(s,a) + \frac{\gamma}{N}\sum_{i_0=1}^{N} U^v_{1,c}\big(\theta_0(s,a), i_0\big), \qquad a \in \mathcal{A}. \tag{49}$$

With $v = V^K$, equation 47–equation 49 are exactly the computations of Algorithm 2, and $\pi_D(s, c) \in \arg\max_a \widetilde{Q}_{V^K}(s, c, a)$.

Lemma 7 (Exact extension identity)**.** *For every $(s, c, a) \in \mathcal{S} \times \Omega^\ell \times \mathcal{A}$,*

$$\widetilde{Q}_{V_N^\star}(s, c, a) = \widehat{Q}^\star(s, c, a). \tag{50}$$

Proof. Fix $c = (\theta_0, \dots, \theta_{\ell-1}) \in \Omega^\ell$ and write $U_m := U_{m,c}^{V_N^\star}$. We first prove, by backward induction on $m \in \{1, \dots, \ell\}$, that for all $u \in \mathcal{S}$ and $i_0, \dots, i_{m-1} \in [N]$,

$$U_m(u, i_0, \dots, i_{m-1}) = \widehat{V}^\star\big(u, \theta_{m:\ell-1}, \Theta^{i_0}, \dots, \Theta^{i_{m-1}}\big). \tag{51}$$

For $m = \ell$, the terminal condition equation 47 and Lemma 6 give $U_\ell(u, i_0, \dots, i_{\ell-1}) = V_N^\star(u, i_0, \dots, i_{\ell-1}) = \widehat{V}^\star(u, \Theta^{i_0}, \dots, \Theta^{i_{\ell-1}})$, which is equation 51 at level ℓ (the list $\theta_{\ell:\ell-1}$ being empty). Suppose equation 51 holds at level $m+1$ for some $m \in \{1, \dots, \ell-1\}$. By equation 48 and the induction hypothesis,

$$\begin{aligned} U_m(u, i_0, \dots, i_{m-1}) &= \max_{a' \in \mathcal{A}} \Big\{ r(u, a') + \frac{\gamma}{N} \sum_{i_m=1}^{N} \widehat{V}^\star\big(\theta_m(u, a'), \theta_{m+1:\ell-1}, \Theta^{i_0}, \dots, \Theta^{i_m}\big) \Big\} \\ &= (\widehat{\mathcal{T}}_\ell \widehat{V}^\star)\big(u, \theta_{m:\ell-1}, \Theta^{i_0}, \dots, \Theta^{i_{m-1}}\big) = \widehat{V}^\star\big(u, \theta_{m:\ell-1}, \Theta^{i_0}, \dots, \Theta^{i_{m-1}}\big), \end{aligned}$$

where the second equality is the definition equation 34 applied to the window $(\theta_m, \theta_{m+1:\ell-1}, \Theta^{i_0}, \dots, \Theta^{i_{m-1}}) \in \Omega^\ell$, whose first table is θ_m and whose appended table is Θ^{i_m}, and the last equality uses $\widehat{\mathcal{T}}_\ell \widehat{V}^\star = \widehat{V}^\star$. This proves equation 51 for every m. Finally, applying equation 51 at level $m = 1$ in equation 49,

$$\widetilde{Q}_{V_N^\star}(s, c, a) = r(s, a) + \frac{\gamma}{N} \sum_{i_0=1}^{N} \widehat{V}^\star\big(\theta_0(s, a), \theta_{1:\ell-1}, \Theta^{i_0}\big) = \widehat{Q}^\star(s, c, a),$$

by the definition equation 36. □

Lemma 8 (Stability of the extension)**.** *For all arrays $v_1, v_2 : \mathcal{D}_N \to \mathbb{R}$ and every level $m \in \{1, \dots, \ell\}$,*

$$\sup_{c, u, i_0, \dots, i_{m-1}} \big| U_{m,c}^{v_1}(u, i_0, \dots, i_{m-1}) - U_{m,c}^{v_2}(u, i_0, \dots, i_{m-1}) \big| \le \gamma^{\ell-m} \|v_1 - v_2\|_\infty. \tag{52}$$

Consequently,

$$\sup_{s, c, a} \big| \widetilde{Q}_{v_1}(s, c, a) - \widetilde{Q}_{v_2}(s, c, a) \big| \le \gamma^\ell \|v_1 - v_2\|_\infty. \tag{53}$$

Proof. Fix c; the recursion equation 48 never modifies it, and all bounds below are uniform in c. We prove equation 52 by backward induction on m. For $m = \ell$, equation 47 gives $U_{\ell,c}^{v_1} - U_{\ell,c}^{v_2} = v_1 - v_2$ on $\mathcal{D}_N$, so equation 52 holds with equality. Assume it at level $m+1$ and fix $u, i_0, \dots, i_{m-1}$. In equation 48, the two braces associated with v_1 and v_2 share the term $r(u, a')$, so for every a' their difference is

$$\frac{\gamma}{N} \sum_{i_m=1}^{N} \big[U_{m+1,c}^{v_1} - U_{m+1,c}^{v_2} \big] \big(\theta_m(u, a'), i_0, \dots, i_m \big),$$

whose absolute value is at most $\gamma \cdot \gamma^{\ell-m-1} \|v_1 - v_2\|_\infty$ by the induction hypothesis. Since maximizing over a' is 1-Lipschitz, the same bound holds for $|U_{m,c}^{v_1} - U_{m,c}^{v_2}|(u, i_0, \dots, i_{m-1})$. For equation 53, the term $r(s, a)$ cancels in equation 49, so $|\widetilde{Q}_{v_1} - \widetilde{Q}_{v_2}|(s, c, a)$ is γ times an average of differences at level 1, hence at most $\gamma \cdot \gamma^{\ell-1} \|v_1 - v_2\|_\infty$. □

Remark 2 (Deployment cost)**.** *Proceeding level by level as in equation 45, computing $U_{m,c}^v$ from $U_{m+1,c}^v$ costs $O(|\mathcal{S}| N^{m+1})$ for the averages over i_m and $O(|\mathcal{S}||\mathcal{A}| N^m)$ for the maxima, and equation 49 costs $O(|\mathcal{S}|N + |\mathcal{A}|)$. The cost of computing all action scores for one queried pair (s, c) is therefore*

$$O\Big(\sum_{m=0}^{\ell-1} \big(|\mathcal{S}| N^{m+1} + |\mathcal{S}||\mathcal{A}| N^m \big) \Big) = O\big(\ell |\mathcal{S}| N^\ell + \ell |\mathcal{S}||\mathcal{A}| N^{\ell-1} \big) \subseteq O\big(\ell |\mathcal{S}| (|\mathcal{A}| + 1) N^\ell \big). \tag{54}$$

The terminal array v, of size $|\mathcal{S}| N^\ell$, is already stored. During a query, the largest temporary array is $U_{\ell-1,c}^v$, with $|\mathcal{S}| N^{\ell-1}$ entries, and $U_{m+1,c}^v$ can be discarded once $U_{m,c}^v$ has been computed. The query thus needs only $O(|\mathcal{S}| N^{\ell-1})$ additional working memory.

Conclusion. Run Algorithm 1 with

$$N := \left\lceil \frac{8}{\varepsilon^2(1-\gamma)^4}\Big(\log\frac{2}{\delta} + \big(1+(\ell-1)|\mathcal{S}||\mathcal{A}|\big)\log|\mathcal{S}|\Big)\right\rceil, \qquad K := \left\lceil \frac{1}{1-\gamma}\log\frac{4}{\varepsilon(1-\gamma)}\right\rceil. \tag{55}$$

It returns $D = (\Theta^1, \dots, \Theta^N, V^K)$, and Algorithm 2 defines the policy π_D, greedy with respect to $\widetilde{Q}_{V^K}$.

Model error. Set

$$\eta := \frac{\varepsilon(1-\gamma)^2 V_{\max}}{4} = \frac{\varepsilon(1-\gamma)}{4}. \tag{56}$$

Since $V_{\max}^2/(2\eta^2) = 8/(\varepsilon^2(1-\gamma)^4)$, the choice of N in equation 55 makes the right-hand side of equation 38 at most δ, so $\mathbb{P}(\mathcal{E}_\eta) \geq 1-\delta$. On $\mathcal{E}_\eta$, Lemma 5 gives

$$\sup_{s,c,a}\big|\widehat{Q}^\star(s,c,a) - Q_\ell^\star(s,c,a)\big| \leq \frac{\gamma\eta}{1-\gamma} \leq \frac{\varepsilon(1-\gamma)V_{\max}}{4}. \tag{57}$$

Computation error. By contraction of $\widehat{\mathcal{T}}_N$ and the choice of K,

$$\|V^K - V_N^\star\|_\infty \leq \gamma^K V_{\max} \leq e^{-(1-\gamma)K} V_{\max} \leq \frac{\varepsilon(1-\gamma)V_{\max}}{4}.$$

Lemma 7 identifies $\widetilde{Q}_{V_N^\star}$ with $\widehat{Q}^\star$, and Lemma 8 then gives

$$\sup_{s,c,a}\big|\widetilde{Q}_{V^K}(s,c,a) - \widehat{Q}^\star(s,c,a)\big| = \sup_{s,c,a}\big|\widetilde{Q}_{V^K}(s,c,a) - \widetilde{Q}_{V_N^\star}(s,c,a)\big| \tag{58}$$

$$\leq \gamma^\ell \|V^K - V_N^\star\|_\infty \tag{59}$$

$$\leq \frac{\varepsilon(1-\gamma)V_{\max}}{4}. \tag{60}$$

This bound is deterministic. Combining equation 57 and equation 60, on $\mathcal{E}_\eta$,

$$\sup_{s,c,a}\big|\widetilde{Q}_{V^K}(s,c,a) - Q_\ell^\star(s,c,a)\big| \leq \frac{\varepsilon(1-\gamma)V_{\max}}{2}. \tag{61}$$

From scores to values. Fix (s,c), let $a^\star$ maximize $Q_\ell^\star(s,c,\cdot)$, and recall that $\pi_D(s,c)$ maximizes $\widetilde{Q}_{V^K}(s,c,\cdot)$. By equation 61,

$$\begin{aligned} V_\ell^\star(s,c) = Q_\ell^\star(s,c,a^\star) &\leq \widetilde{Q}_{V^K}(s,c,a^\star) + \frac{\varepsilon(1-\gamma)V_{\max}}{2} \\ &\leq \widetilde{Q}_{V^K}(s,c,\pi_D(s,c)) + \frac{\varepsilon(1-\gamma)V_{\max}}{2} \\ &\leq Q_\ell^\star(s,c,\pi_D(s,c)) + \varepsilon(1-\gamma)V_{\max}. \end{aligned}$$

Let $\mathcal{T}_\ell^{\pi_D}$ denote the Bellman evaluation operator of π_D in the true augmented MDP. Since $Q_\ell^\star(s,c,\pi_D(s,c)) = (\mathcal{T}_\ell^{\pi_D} V_\ell^\star)(s,c)$, the preceding inequality reads $V_\ell^\star \leq \mathcal{T}_\ell^{\pi_D} V_\ell^\star + \varepsilon(1-\gamma)V_{\max}\mathbf{1}$. As $V_\ell^{\pi_D} = \mathcal{T}_\ell^{\pi_D} V_\ell^{\pi_D}$ and $\mathcal{T}_\ell^{\pi_D}$ is a monotone γ-contraction,

$$\|V_\ell^\star - V_\ell^{\pi_D}\|_\infty \leq \gamma\|V_\ell^\star - V_\ell^{\pi_D}\|_\infty + \varepsilon(1-\gamma)V_{\max}, \qquad \text{hence} \qquad \|V_\ell^\star - V_\ell^{\pi_D}\|_\infty \leq \varepsilon V_{\max}.$$

This proves equation 7 simultaneously for every $(s,c) \in \mathcal{S} \times \Omega^\ell$, on the event $\mathcal{E}_\eta$ of probability at least $1-\delta$.

Computational guarantees. Sampling the dictionary costs $O(N|\mathcal{S}||\mathcal{A}|)$ and, by Remark 1, the K value-iteration steps cost $O(K|\mathcal{S}|(|\mathcal{A}|+1)N^\ell)$, so preprocessing costs $O\big(N|\mathcal{S}||\mathcal{A}| + K|\mathcal{S}|(|\mathcal{A}|+1)N^\ell\big)$. The representation $D = (\Theta^1, \dots, \Theta^N, V^K)$ occupies $O(N|\mathcal{S}||\mathcal{A}| + |\mathcal{S}|N^\ell)$ memory locations. By Remark 2, selecting an action at a queried pair (s,c) costs $O(\ell|\mathcal{S}|(|\mathcal{A}|+1)N^\ell)$ time and $O(|\mathcal{S}|N^{\ell-1})$ working memory, the observed window itself occupying $O(\ell|\mathcal{S}||\mathcal{A}|)$; the total memory during online action selection is therefore $O\big(N|\mathcal{S}||\mathcal{A}| + |\mathcal{S}|N^\ell + \ell|\mathcal{S}||\mathcal{A}|\big)$. Finally, by equation 55,

$$N = O\left(\frac{\log(1/\delta) + \big(1+(\ell-1)|\mathcal{S}||\mathcal{A}|\big)\log|\mathcal{S}|}{\varepsilon^2(1-\gamma)^4}\right), \qquad K = O\left(\frac{\log\big(1/[\varepsilon(1-\gamma)]\big)}{1-\gamma}\right),$$

so for every fixed ℓ, preprocessing time, representation size, and per-decision computation are polynomial in $|\mathcal{S}|$, $|\mathcal{A}|$, ε^{-1}, $\log(1/\delta)$, and $(1-\gamma)^{-1}$. This completes the proof. □

A.3 PROOF OF THEOREM 3

Throughout this section, M denotes the number of observed tables used at a planner update, $\Theta_0, \dots, \Theta_{M-1}$ these tables, and $\widehat{P}_M$, $\widehat{L}_M = \bigotimes_{(s,a)} \widehat{P}_M(\cdot \mid s, a)$ the empirical model of Section 5, equation equation 11. Note that $\widehat{L}_M$ is a *product* law, not the empirical distribution of the dictionary used in Appendix A.2. We write $V_{\max} := (1-\gamma)^{-1}$ and use the shorthand $\theta_{k:k'} = (\theta_k, \dots, \theta_{k'})$.

Let $\widehat{V}^\star_M$ and $\widehat{Q}^\star_M$ denote the optimal value and action-value functions of the augmented MDP on $\mathcal{S} \times \Omega^\ell$ when each incoming table has law $\widehat{L}_M$. At an update, RPTAS-OFFLINE draws

$$Z^1, \dots, Z^{N_M} \overset{\text{i.i.d.}}{\sim} \widehat{L}_M, \qquad N_M := \lceil M V_{\max}^3 \rceil,$$

runs $K_M := \lceil V_{\max} \log(V_{\max} M^2) \rceil$ value-iteration steps from zero on $\mathcal{S} \times [N_M]^\ell$, and returns $D_M = (Z^1, \dots, Z^{N_M}, v^{K_M})$. Given a queried pair (s, c), RPTAS-ONLINE returns the action scores $\widetilde{Q}_M(s, c, \cdot)$ obtained from equation 48–equation 49 with dictionary $Z^1, \dots, Z^{N_M}$ and terminal array v^{K_M}. The deployed policy π_M is greedy with respect to $\widetilde{Q}_M$ with a fixed tie-breaking rule; it is a deterministic stationary policy on the augmented state space. At time 0, the algorithm executes a fixed action A_0; we let π_0 denote the constant policy that always plays A_0.

For $t \geq 1$, let $M_t := 2^{\lfloor \log_2 t \rfloor}$ be the number of observed tables at the most recent update before time t, and set $W_t := V_\ell^{\pi_{M_t}}$ for $t \geq 1$, $W_0 := V_\ell^{\pi_0}$. Thus W_t is the true infinite-horizon value of the policy deployed at time t, and is constant as a function between two updates. The epochs are $\{0\}$ and the nonempty sets $\{M, \dots, \min(2M-1, T-1)\}$ for dyadic $M < T$; their number is at most

$$J := 2 + \lceil \log_2 T \rceil.$$

We abbreviate the two logarithmic factors that appear below as

$$\Lambda := \log \frac{16 J |\mathcal{S}| |\Omega|^{\ell-1}}{\delta}, \qquad \Lambda_0 := |\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)\log 2 + \log \frac{16J}{\delta}, \tag{62}$$

and define the burn-in size

$$M_0 := \lceil 16 V_{\max}^2 \Lambda_0 \rceil. \tag{63}$$

Since $|\Omega| = |\mathcal{S}|^{|\mathcal{S}||\mathcal{A}|}$, we have $\Lambda = \log(16J/\delta) + \big(1 + (\ell-1)|\mathcal{S}||\mathcal{A}|\big)\log|\mathcal{S}|$.

Finally, for a deterministic stationary policy π and an augmented state x, we write $\mathbb{E}^\pi_x$ for the expectation along the auxiliary augmented chain $(\bar{X}_h)_{h \geq 0}$ with $\bar{X}_0 = x$, actions chosen by π, and each incoming table drawn independently from the true law L. We write $\bar{X}_h = (\bar{S}_h, \bar{\Theta}_{h:h+\ell-1})$, so that the tables $\bar{\Theta}_{0:\ell-1}$ are those of x and $\bar{\Theta}_{h+\ell} \sim L$ is the table appended at step h. This chain is a device for the analysis and is distinct from the trajectory $(X_t)_{t \geq 0}$ generated by the algorithm.

Adding and subtracting $(1-\gamma)W_t(X_t)$ in the definition of regret gives

$$\begin{aligned} \mathrm{Reg}(T) &= \sum_{t=0}^{T-1} \big[(1-\gamma)V_\ell^\star(X_t) - r_t\big] \\ &= \underbrace{(1-\gamma)\sum_{t=0}^{T-1} \big[V_\ell^\star(X_t) - W_t(X_t)\big]}_{\text{policy loss}} + \underbrace{\sum_{t=0}^{T-1} \big[(1-\gamma)W_t(X_t) - r_t\big]}_{\text{value-to-reward error}}. \end{aligned} \tag{64}$$

We bound the two terms in turn. The policy loss is controlled by Lemmas 9–13, the value-to-reward error by Lemma 14.

Fix an update based on M observed tables. Adding and subtracting $\widehat{V}^\star_M$ gives

$$V_\ell^\star - V_\ell^{\pi_M} = (V_\ell^\star - \widehat{V}^\star_M) + (\widehat{V}^\star_M - V_\ell^{\pi_M}). \tag{65}$$

For $x = (s, \theta_{0:\ell-1})$, the Bellman equations give

$$\begin{aligned} \widehat{Q}^\star_M(x, \pi_M(x)) &= r(s, \pi_M(x)) + \gamma\, \mathbb{E}_{\Theta \sim \widehat{L}_M}\big[\widehat{V}^\star_M\big(\theta_0(s, \pi_M(x)), \theta_{1:\ell-1}, \Theta\big)\big], \\ V_\ell^{\pi_M}(x) &= r(s, \pi_M(x)) + \gamma\, \mathbb{E}_{\Theta \sim L}\big[V_\ell^{\pi_M}\big(\theta_0(s, \pi_M(x)), \theta_{1:\ell-1}, \Theta\big)\big]. \end{aligned}$$

Subtracting the second equality from the first, adding $\widehat{V}^\star_M(x) - \widehat{Q}^\star_M(x, \pi_M(x))$ to both sides, adding and subtracting $\gamma\, \mathbb{E}_{\Theta\sim L}[\widehat{V}^\star_M(\cdot)]$, and writing $\widehat{V}^\star_M = V^\star_\ell + (\widehat{V}^\star_M - V^\star_\ell)$ in the difference of expectations yields

$$\begin{aligned}\widehat{V}^\star_M(x) - V^{\pi_M}_\ell(x) = \widehat{V}^\star_M(x) - \widehat{Q}^\star_M(x, \pi_M(x)) \\ + \gamma\, \mathbb{E}_{\Theta\sim L}\big[(\widehat{V}^\star_M - V^{\pi_M}_\ell)\big(\theta_0(s, \pi_M(x)), \theta_{1:\ell-1}, \Theta\big)\big] \\ + \gamma\big(\mathbb{E}_{\Theta\sim \widehat{L}_M} - \mathbb{E}_{\Theta\sim L}\big)\big[V^\star_\ell\big(\theta_0(s, \pi_M(x)), \theta_{1:\ell-1}, \Theta\big)\big] \\ + \gamma\big(\mathbb{E}_{\Theta\sim \widehat{L}_M} - \mathbb{E}_{\Theta\sim L}\big)\big[(\widehat{V}^\star_M - V^\star_\ell)\big(\theta_0(s, \pi_M(x)), \theta_{1:\ell-1}, \Theta\big)\big]. \qquad (66)\end{aligned}$$

Fix the observed tables and the planner's randomization, so that $\widehat{L}_M$, $\widehat{V}^\star_M$ and π_M are fixed. Repeatedly substituting equation 66 into its own continuation term along the auxiliary chain $(\bar{X}_h)$ under $\mathbb{E}^{\pi_M}_x$, and applying the tower property, gives for every $H \geq 1$,

$$\begin{aligned}\widehat{V}^\star_M(x) - V^{\pi_M}_\ell(x) = \sum_{h=0}^{H-1} \gamma^h\, \mathbb{E}^{\pi_M}_x\big[\widehat{V}^\star_M(\bar{X}_h) - \widehat{Q}^\star_M(\bar{X}_h, \pi_M(\bar{X}_h))\big] \\ + \sum_{h=0}^{H-1} \gamma^{h+1}\, \mathbb{E}^{\pi_M}_x\Big[\big(\mathbb{E}_{\Theta\sim \widehat{L}_M} - \mathbb{E}_{\Theta\sim L}\big)\big[V^\star_\ell\big(\bar{\Theta}_h(\bar{S}_h, \pi_M(\bar{X}_h)), \bar{\Theta}_{h+1:h+\ell-1}, \Theta\big)\big]\Big] \\ + \sum_{h=0}^{H-1} \gamma^{h+1}\, \mathbb{E}^{\pi_M}_x\Big[\big(\mathbb{E}_{\Theta\sim \widehat{L}_M} - \mathbb{E}_{\Theta\sim L}\big)\big[(\widehat{V}^\star_M - V^\star_\ell)\big(\bar{\Theta}_h(\bar{S}_h, \pi_M(\bar{X}_h)), \bar{\Theta}_{h+1:h+\ell-1}, \Theta\big)\big]\Big] \\ + \gamma^H\, \mathbb{E}^{\pi_M}_x\big[(\widehat{V}^\star_M - V^{\pi_M}_\ell)(\bar{X}_H)\big]. \qquad (67)\end{aligned}$$

The last term is bounded by $\gamma^H V_{\max}$ in absolute value and vanishes as $H \to \infty$. For the first sum, π_M is greedy with respect to $\widetilde{Q}_M$, so at every augmented state (see the proof of equation 86 below for the two-line argument)

$$0 \leq \widehat{V}^\star_M(x) - \widehat{Q}^\star_M(x, \pi_M(x)) \leq 2\|\widetilde{Q}_M - \widehat{Q}^\star_M\|_\infty. \qquad (68)$$

For the third sum, at every augmented state,

$$\Big|\big(\mathbb{E}_{\Theta\sim \widehat{L}_M} - \mathbb{E}_{\Theta\sim L}\big)\big[(\widehat{V}^\star_M - V^\star_\ell)\big(\theta_0(s, \pi_M(x)), \theta_{1:\ell-1}, \Theta\big)\big]\Big| \leq \|\widehat{L}_M - L\|_1 \|\widehat{V}^\star_M - V^\star_\ell\|_\infty.$$

Combining these bounds with equation 65, $\sum_{h\geq 0} \gamma^h = V_{\max}$, and equation 67 as $H \to \infty$ yields

$$\begin{aligned}\|V^\star_\ell - V^{\pi_M}_\ell\|_\infty \leq \underbrace{\|V^\star_\ell - \widehat{V}^\star_M\|_\infty}_{\text{empirical value error}} \\ \underbrace{+ \sup_x \Big|\sum_{h=0}^{\infty} \gamma^{h+1}\, \mathbb{E}^{\pi_M}_x\Big[\big(\mathbb{E}_{\Theta\sim \widehat{L}_M} - \mathbb{E}_{\Theta\sim L}\big)\big[V^\star_\ell\big(\bar{\Theta}_h(\bar{S}_h, \pi_M(\bar{X}_h)), \bar{\Theta}_{h+1:h+\ell-1}, \Theta\big)\big]\Big]\Big|}_{\text{fixed-function transition error}} \\ + \underbrace{\gamma V_{\max}\|\widehat{L}_M - L\|_1 \|\widehat{V}^\star_M - V^\star_\ell\|_\infty}_{\text{data-dependent transition error}} + \underbrace{2V_{\max}\|\widetilde{Q}_M - \widehat{Q}^\star_M\|_\infty}_{\text{finite-planner error}}. \qquad (69)\end{aligned}$$

We first control the fixed-function transition error. The difficulty is that $\widehat{L}_M$ is a product of empirical row distributions, not the empirical distribution of M complete tables. The following lemma shows that, for a fixed function, its error nevertheless obeys a Bernstein inequality with the variance under L.

Lemma 9 (Product-estimator Bernstein inequality)**.** *For any deterministic $f : \Omega \to [0, V_{\max}]$, any $M \geq 1$, and any $w > 0$,*

$$\mathbb{P}\left(\big|\mathbb{E}_{\widehat{L}_M}[f] - \mathbb{E}_L[f]\big| > \sqrt{\frac{2\,\mathrm{Var}_L(f)\,w}{M}} + \frac{2V_{\max}w}{3M}\right) \leq 2e^{-w}. \qquad (70)$$

Proof. For each (s,a), draw a uniform random permutation $\sigma_{s,a}$ of $\{0,\dots,M-1\}$, independently across (s,a) and independently of the observed tables; write $\mathbb{E}_\sigma$ for the expectation over these permutations. For $j \in \{0,\dots,M-1\}$, form the table

$$\Theta^{(j)} := \big(\Theta_{\sigma_{s,a}(j)}(s,a)\big)_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in \Omega.$$

Conditional on $\Theta_0,\dots,\Theta_{M-1}$, for fixed j the entries of $\Theta^{(j)}$ are independent across (s,a), each with distribution $\widehat{P}_M(\cdot \mid s,a)$; hence $\Theta^{(j)}$ has conditional law $\widehat{L}_M$, and averaging over j gives

$$\mathbb{E}_\sigma\left[\frac{1}{M}\sum_{j=0}^{M-1} f\big(\Theta^{(j)}\big)\,\middle|\,\Theta_0,\dots,\Theta_{M-1}\right] = \mathbb{E}_{\widehat{L}_M}[f]. \tag{71}$$

Now fix the permutations. For each (s,a) the indices $\sigma_{s,a}(0),\dots,\sigma_{s,a}(M-1)$ are distinct, so the tables $\Theta^{(0)},\dots,\Theta^{(M-1)}$ use distinct observations from every row. Independence across dates and rows implies that, for fixed permutations, $\Theta^{(0)},\dots,\Theta^{(M-1)}$ are i.i.d. with law L. Conditional Jensen's inequality applied to equation 71 gives, for every $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\exp\big(\lambda\big[\mathbb{E}_{\widehat{L}_M}[f] - \mathbb{E}_L[f]\big]\big) &\le \mathbb{E}\,\mathbb{E}_\sigma \exp\left(\frac{\lambda}{M}\sum_{j=0}^{M-1}\big[f\big(\Theta^{(j)}\big) - \mathbb{E}_L[f]\big]\right)\\ &= \mathbb{E}\exp\left(\frac{\lambda}{M}\sum_{j=0}^{M-1}\big[f(\Theta_j) - \mathbb{E}_L[f]\big]\right).\end{aligned}$$

Thus the moment-generating function of the product-estimator error is bounded by that of the average of M i.i.d. copies of $f(\Theta) - \mathbb{E}_L[f]$, which take values in an interval of length $V_{\max}$. Bernstein's inequality applied to the positive and negative tails proves equation 70. □

Lemma 10 (Model confidence bounds)**.** *There is an event $\mathcal{E}_{\text{model}}$ of probability at least $1 - 3\delta/16$ on which the following three inequalities hold simultaneously for every dyadic update $M < T$.*

1. *For every $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $\theta_{0:\ell-1} \in \Omega^\ell$,*

$$\begin{aligned}&\Big|\mathbb{E}_{\Theta\sim\widehat{L}_M}\big[V_\ell^\star\big(\theta_0(s,a),\theta_{1:\ell-1},\Theta\big)\big] - \mathbb{E}_{\Theta\sim L}\big[V_\ell^\star\big(\theta_0(s,a),\theta_{1:\ell-1},\Theta\big)\big]\Big|\\ &\quad\le \sqrt{\frac{2\Lambda\,\mathrm{Var}_{\Theta\sim L}\big(V_\ell^\star(\theta_0(s,a),\theta_{1:\ell-1},\Theta)\big)}{M}} + \frac{2V_{\max}\Lambda}{3M}.\end{aligned} \tag{72}$$

2. *The laws of transition tables satisfy*

$$\|\widehat{L}_M - L\|_1 \le 2\sqrt{\frac{\Lambda_0}{M}}. \tag{73}$$

3. *Every bounded $g : \mathcal{S}\times\Omega^\ell \to \mathbb{R}$, including one chosen after observing the tables, satisfies*

$$\sup_{s\in\mathcal{S},\ a\in\mathcal{A},\ \theta_{0:\ell-1}\in\Omega^\ell}\Big|\big(\mathbb{E}_{\Theta\sim\widehat{L}_M} - \mathbb{E}_{\Theta\sim L}\big)\big[g\big(\theta_0(s,a),\theta_{1:\ell-1},\Theta\big)\big]\Big| \le 2\sqrt{\frac{\Lambda_0}{M}}\,\|g\|_\infty. \tag{74}$$

Proof. (i). For fixed $u \in \mathcal{S}$ and $\theta_{1:\ell-1} \in \Omega^{\ell-1}$, the function $\Theta \mapsto V_\ell^\star(u,\theta_{1:\ell-1},\Theta)$ is deterministic and does not depend on the observed tables. Apply Lemma 9 to it with $w = \Lambda$. There are at most $|\mathcal{S}||\Omega|^{\ell-1}$ such functions and at most J update sizes, so a union bound gives failure probability at most $2e^{-\Lambda}\,|\mathcal{S}||\Omega|^{\ell-1}J = \delta/8$. Taking $u = \theta_0(s,a)$ yields equation 72.

(ii). By Agrawal (2020, Theorem I.3), applied with exponential parameter $M/2$, each empirical transition row satisfies

$$\mathbb{E}\exp\Big[\tfrac{M}{2}\,\mathrm{KL}\big(\widehat{P}_M(\cdot \mid s,a)\,\big\|\,P(\cdot \mid s,a)\big)\Big] \le 2^{|\mathcal{S}|-1}.$$

The row estimates are independent and the KL divergence is additive over product distributions, hence

$$\mathbb{E}\exp\Big[\tfrac{M}{2}\,\mathrm{KL}(\widehat{L}_M\|L)\Big] \le 2^{|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)}.$$

Markov's inequality gives

$$\mathbb{P}\Big(\mathrm{KL}(\widehat{L}_M\|L) > \tfrac{2\Lambda_0}{M}\Big) \le 2^{|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)}e^{-\Lambda_0} = \frac{\delta}{16J}.$$

A union bound over the at most J updates shows that this KL bound holds at every update with probability at least $1-\delta/16$, and Pinsker's inequality then yields $\|\widehat{L}_M - L\|_1 \le \sqrt{2\,\mathrm{KL}(\widehat{L}_M\|L)} \le 2\sqrt{\Lambda_0/M}$.

(iii). Fix the observed tables, a bounded g, and $(s,a,\theta_{0:\ell-1})$. The difference of expectations in equation 74 equals $\sum_{\Theta\in\Omega}\big(\widehat{L}_M(\Theta) - L(\Theta)\big)g(\theta_0(s,a),\theta_{1:\ell-1},\Theta)$, whose absolute value is at most $\|\widehat{L}_M - L\|_1\|g\|_\infty$. This holds for all g and all $(s,a,\theta_{0:\ell-1})$ simultaneously, so (iii) is a deterministic consequence of (ii).

The total failure probability is $\delta/8 + \delta/16 = 3\delta/16$. □

The fixed-function term in equation 69 accumulates the one-step errors equation 72 along a discounted trajectory. To retain the variance dependence of the Bernstein bound, we control the discounted sum of conditional standard deviations of $V_\ell^\star$.

Lemma 11 (Discounted variance bound)**.** *Let π be a deterministic stationary policy. For every augmented state x,*

$$\sum_{h=0}^{\infty}\gamma^{h+1}\,\mathbb{E}_x^\pi\Big[\sqrt{\mathrm{Var}\big(V_\ell^\star(\bar{X}_{h+1})\mid \bar{X}_h\big)}\Big] \le \sqrt{2V_{\max}^3}, \tag{75}$$

where the conditional variance is over the appended table $\bar{\Theta}_{h+\ell}\sim L$ given $\bar{X}_h$. Consequently, on $\mathcal{E}_{\mathrm{model}}$, every deterministic stationary policy π, including one selected using the observed tables, satisfies

$$\sup_x\Big|\sum_{h=0}^{\infty}\gamma^{h+1}\,\mathbb{E}_x^\pi\Big[\big(\mathbb{E}_{\Theta\sim\widehat{L}_M} - \mathbb{E}_{\Theta\sim L}\big)\big[V_\ell^\star\big(\bar{\Theta}_h(\bar{S}_h,\pi(\bar{X}_h)),\bar{\Theta}_{h+1:h+\ell-1},\Theta\big)\big]\Big]\Big| \tag{76}$$

$$\le 2\sqrt{\frac{V_{\max}^3\Lambda}{M}} + \frac{2V_{\max}^2\Lambda}{3M}. \tag{77}$$

For $\ell = 1$, the list $\bar{\Theta}_{h+1:h+\ell-1}$ is empty.

Proof. Proof of equation 75. Fix x. Bellman optimality and nonnegative rewards imply, for every augmented state y,

$$V_\ell^\star(y) - \gamma\,\mathbb{E}_y^\pi\big[V_\ell^\star(\bar{X}_1)\big] \ge r(y,\pi(y)) \ge 0. \tag{78}$$

Also $0 \le V_\ell^\star \le V_{\max}$. Expanding the conditional variance,

$$\begin{aligned}\gamma^2\,\mathrm{Var}\big(V_\ell^\star(\bar{X}_1)\mid \bar{X}_0 = y\big) &= \gamma^2\mathbb{E}_y^\pi\big[V_\ell^\star(\bar{X}_1)^2\big] - \big(\gamma\,\mathbb{E}_y^\pi[V_\ell^\star(\bar{X}_1)]\big)^2\\ &\le \gamma\,\mathbb{E}_y^\pi\big[V_\ell^\star(\bar{X}_1)^2\big] - V_\ell^\star(y)^2 + 2V_{\max}\big(V_\ell^\star(y) - \gamma\,\mathbb{E}_y^\pi[V_\ell^\star(\bar{X}_1)]\big).\end{aligned} \tag{79}$$

Here $\gamma^2 \le \gamma$ bounds the first term; for the squared mean, write $\gamma\,\mathbb{E}_y^\pi[V_\ell^\star(\bar{X}_1)]$ as $V_\ell^\star(y)$ minus the nonnegative quantity in equation 78, expand the square, and drop its nonpositive quadratic term.

Apply equation 79 at $y = \bar{X}_h$, multiply by γ^h, take $\mathbb{E}_x^\pi$, and sum over $h \ge 0$. Both terms on the right telescope, and the terminal terms vanish because $V_\ell^\star$ is bounded and $\gamma < 1$:

$$\sum_{h=0}^{\infty}\gamma^h\,\mathbb{E}_x^\pi\big[\gamma^2\,\mathrm{Var}\big(V_\ell^\star(\bar{X}_{h+1})\mid \bar{X}_h\big)\big] \le -V_\ell^\star(x)^2 + 2V_{\max}V_\ell^\star(x) \le 2V_{\max}^2.$$

Jensen's inequality ($\mathbb{E}\sqrt{\cdot} \le \sqrt{\mathbb{E}\,\cdot}$), Cauchy–Schwarz, and $\sum_{h\ge 0}\gamma^h = V_{\max}$ give

$$\begin{aligned}
&\sum_{h=0}^{\infty} \gamma^{h+1}\, \mathbb{E}_x^{\pi}\Big[\sqrt{\mathrm{Var}\big(V_\ell^\star(\bar{X}_{h+1}) \mid \bar{X}_h\big)}\Big] \\
&\le \sqrt{\Big(\sum_{h=0}^{\infty}\gamma^h\Big)\Big(\sum_{h=0}^{\infty}\gamma^h\, \mathbb{E}_x^{\pi}\big[\gamma^2\,\mathrm{Var}\big(V_\ell^\star(\bar{X}_{h+1}) \mid \bar{X}_h\big)\big]\Big)} \\
&\le \sqrt{2V_{\max}^3}.
\end{aligned}$$

Proof of equation 77. Conditional on $\bar{X}_h$, the successor $\bar{\Theta}_h(\bar{S}_h, \pi(\bar{X}_h))$ and the tables $\bar{\Theta}_{h+1:h+\ell-1}$ are fixed; only the appended table is averaged over L or $\widehat{L}_M$. Hence equation 72 bounds the difference of the two expectations by

$$\sqrt{\frac{2\Lambda}{M}}\,\sqrt{\mathrm{Var}\big(V_\ell^\star(\bar{X}_{h+1}) \mid \bar{X}_h\big)} + \frac{2V_{\max}\Lambda}{3M}.$$

Multiply by γ^{h+1}, take $\mathbb{E}_x^{\pi}$, and sum over h. The variance terms are bounded by equation 75, which gives $2\sqrt{V_{\max}^3\Lambda/M}$; the constant terms sum to at most $\frac{2V_{\max}\Lambda}{3M}\sum_{h\ge 0}\gamma^{h+1} \le \frac{2V_{\max}^2\Lambda}{3M}$. The argument applies to a data-dependent policy because equation 72 holds simultaneously at every augmented state and action; no union bound over policies is needed. □

The remaining statistical quantity in equation 69 is $\|\widehat{V}_M^\star - V_\ell^\star\|_\infty$. Comparing the empirical and true Bellman optimality equations yields an inequality in which this quantity appears on both sides, with coefficient $\gamma V_{\max}\|\widehat{L}_M - L\|_1$ on the right. For $M \ge M_0$ this coefficient is at most $1/2$ and the term can be absorbed. The next lemma makes this explicit and also bounds the loss of any policy with a small empirical Bellman residual.

Lemma 12 (Stability)**.** *Suppose $M \ge M_0$. On $\mathcal{E}_{\mathrm{model}}$,*

$$\|\widehat{V}_M^\star - V_\ell^\star\|_\infty \le 4\sqrt{\frac{V_{\max}^3\Lambda}{M}} + \frac{4V_{\max}^2\Lambda}{3M}. \tag{80}$$

Moreover, if a deterministic stationary policy π has empirical Bellman residual at most $b \ge 0$, i.e. for every $x = (s, \theta_{0:\ell-1})$,

$$0 \le \widehat{V}_M^\star(x) - r(s,\pi(x)) - \gamma\,\mathbb{E}_{\Theta\sim\widehat{L}_M}\big[\widehat{V}_M^\star\big(\theta_0(s,\pi(x)),\theta_{1:\ell-1},\Theta\big)\big] \le b, \tag{81}$$

then its true value satisfies

$$\|V_\ell^\star - V_\ell^\pi\|_\infty \le 8\sqrt{\frac{V_{\max}^3\Lambda}{M}} + \frac{8V_{\max}^2\Lambda}{3M} + V_{\max} b. \tag{82}$$

Proof. By equation 73 and $M \ge M_0 \ge 16V_{\max}^2\Lambda_0$,

$$\gamma V_{\max}\|\widehat{L}_M - L\|_1 \le 2V_{\max}\sqrt{\Lambda_0/M} \le \tfrac{1}{2}. \tag{83}$$

Proof of equation 80. Let $\widehat{\pi}_M^\star$ be an optimal policy for the empirical model. Fix $x = (s, \theta_{0:\ell-1})$. The empirical Bellman equation for $\widehat{\pi}_M^\star$ and the true Bellman inequality for the same action give

$$\begin{aligned}
\widehat{V}_M^\star(x) - V_\ell^\star(x) \le\ & \gamma\,\mathbb{E}_{\Theta\sim L}\big[(\widehat{V}_M^\star - V_\ell^\star)\big(\theta_0(s,\widehat{\pi}_M^\star(x)),\theta_{1:\ell-1},\Theta\big)\big] \\
&+ \gamma\big(\mathbb{E}_{\Theta\sim\widehat{L}_M} - \mathbb{E}_{\Theta\sim L}\big)\big[V_\ell^\star\big(\theta_0(s,\widehat{\pi}_M^\star(x)),\theta_{1:\ell-1},\Theta\big)\big] \\
&+ \gamma\big(\mathbb{E}_{\Theta\sim\widehat{L}_M} - \mathbb{E}_{\Theta\sim L}\big)\big[(\widehat{V}_M^\star - V_\ell^\star)\big(\theta_0(s,\widehat{\pi}_M^\star(x)),\theta_{1:\ell-1},\Theta\big)\big].
\end{aligned}$$

Iterate this inequality along the auxiliary chain under $\mathbb{E}_x^{\widehat{\pi}_M^\star}$. The discounted sum of the second term is bounded by equation 77. For the third term, equation 74 gives at every state the bound $\|\widehat{L}_M - L\|_1\|\widehat{V}_M^\star - V_\ell^\star\|_\infty$, whose discounted sum is at most $\gamma V_{\max}\|\widehat{L}_M - L\|_1\|\widehat{V}_M^\star - V_\ell^\star\|_\infty$.

For the opposite difference $V_\ell^\star - \widehat{V}_M^\star$, repeat the argument with a true optimal policy $\pi_\ell^\star$: the true Bellman equation holds for this policy and the empirical Bellman optimality inequality holds for its action; the terms involving $V_\ell^\star$ change sign but their absolute values obey the same bounds. Combining the two directions,

$$\|\widehat{V}_M^\star - V_\ell^\star\|_\infty \le 2\sqrt{\frac{V_{\max}^3\Lambda}{M}} + \frac{2V_{\max}^2\Lambda}{3M} + \gamma V_{\max}\|\widehat{L}_M - L\|_1\|\widehat{V}_M^\star - V_\ell^\star\|_\infty. \tag{84}$$

By equation 83, the last term is at most $\frac{1}{2}\|\widehat{V}_M^\star - V_\ell^\star\|_\infty$; moving it to the left proves equation 80.

Proof of equation 82. Subtract the true Bellman equation of π from equation 81 and iterate along the auxiliary chain under $\mathbb{E}_x^\pi$. The discounted sum of the empirical residual is at most $V_{\max}b$. Splitting $\widehat{V}_M^\star = V_\ell^\star + (\widehat{V}_M^\star - V_\ell^\star)$ in the remaining model error gives two terms: by equation 77, the part involving $V_\ell^\star$ contributes at most $2\sqrt{V_{\max}^3\Lambda/M} + 2V_{\max}^2\Lambda/(3M)$; by equation 74 and equation 83, the other part contributes at most $\frac{1}{2}\|\widehat{V}_M^\star - V_\ell^\star\|_\infty$. Hence

$$\|\widehat{V}_M^\star - V_\ell^\pi\|_\infty \le V_{\max}b + 2\sqrt{\frac{V_{\max}^3\Lambda}{M}} + \frac{2V_{\max}^2\Lambda}{3M} + \frac{1}{2}\|\widehat{V}_M^\star - V_\ell^\star\|_\infty.$$

Finally,

$$\begin{aligned}
&\|V_\ell^\star - V_\ell^\pi\|_\infty \\
&\le \|V_\ell^\star - \widehat{V}_M^\star\|_\infty + \|\widehat{V}_M^\star - V_\ell^\pi\|_\infty \\
&\le \tfrac{3}{2}\|\widehat{V}_M^\star - V_\ell^\star\|_\infty + V_{\max}b + 2\sqrt{V_{\max}^3\Lambda/M} + 2V_{\max}^2\Lambda/(3M)
\end{aligned}$$

, and equation 80 gives equation 82. □

It remains to control the finite-planner error $2V_{\max}\|\widetilde{Q}_M - \widehat{Q}_M^\star\|_\infty$ in equation 69. Conditionally on the observed tables, $\widehat{V}_M^\star$ is fixed and the dictionary entries $Z^1, \dots, Z^{N_M}$ are independent draws from $\widehat{L}_M$. We bound the dictionary sampling error by conditional concentration, then account for the finite number of value-iteration steps, exactly as in the known-model analysis of Appendix A.2.

Lemma 13 (Uniform action-score accuracy)**.** *There is an event $\mathcal{E}_{\mathrm{dict}}$ of probability at least $1 - \delta/8$ on which, at every dyadic update $M < T$,*

$$\|\widetilde{Q}_M - \widehat{Q}_M^\star\|_\infty \le \sqrt{\frac{V_{\max}\Lambda}{2M}} + M^{-2}, \tag{85}$$

where the norm ranges over all augmented state–action pairs, including windows outside the sampled dictionary. Consequently, on $\mathcal{E}_{\mathrm{model}} \cap \mathcal{E}_{\mathrm{dict}}$, every update $M \ge M_0$ satisfies

$$\|V_\ell^\star - V_\ell^{\pi_M}\|_\infty \le 8\sqrt{\frac{V_{\max}^3\Lambda}{M}} + \frac{8V_{\max}^2\Lambda}{3M} + 2V_{\max}\left[\sqrt{\frac{V_{\max}\Lambda}{2M}} + M^{-2}\right]. \tag{86}$$

Proof. Proof of equation 85. Fix an update and condition on the M observed tables; then $\widehat{L}_M$ and $\widehat{V}_M^\star$ are fixed and $Z^1, \dots, Z^{N_M}$ are i.i.d. with law $\widehat{L}_M$. For each $u \in \mathcal{S}$ and $\theta_{1:\ell-1} \in \Omega^{\ell-1}$, Hoeffding's inequality applies to the function $Z \mapsto \widehat{V}_M^\star(u, \theta_{1:\ell-1}, Z) \in [0, V_{\max}]$. There are at most $|\mathcal{S}||\Omega|^{\ell-1}$ such functions, so a union bound shows that, except with conditional probability at most $\delta/(8J)$,

$$\sup_{u,\theta_{1:\ell-1}} \left| \frac{1}{N_M}\sum_{j=1}^{N_M} \widehat{V}_M^\star(u, \theta_{1:\ell-1}, Z^j) - \mathbb{E}_{\Theta\sim\widehat{L}_M}\left[\widehat{V}_M^\star(u, \theta_{1:\ell-1}, \Theta)\right] \right| \le V_{\max}\sqrt{\frac{\Lambda}{2N_M}}. \tag{87}$$

For this proof only, let V_M^{dict} and Q_M^{dict} be the exact optimal value and action scores of the augmented MDP when each incoming table is drawn uniformly from $\{Z^1, \dots, Z^{N_M}\}$. The Bellman operators

for this law and for $\widehat{L}_M$ are γ-contractions; comparing them at $\widehat{V}^\star_M$ using equation 87 gives

$$\|V_M^{\mathrm{dict}} - \widehat{V}^\star_M\|_\infty \le \gamma V_{\max}^2 \sqrt{\frac{\Lambda}{2N_M}},$$
$$\|Q_M^{\mathrm{dict}} - \widehat{Q}^\star_M\|_\infty \le \gamma \|V_M^{\mathrm{dict}} - \widehat{V}^\star_M\|_\infty + \gamma V_{\max}\sqrt{\frac{\Lambda}{2N_M}}$$
$$\le \gamma V_{\max}^2 \sqrt{\frac{\Lambda}{2N_M}},$$

where the last step uses $1 + \gamma V_{\max} = V_{\max}$.

On dictionary windows, the Bellman equation of V_M^{dict} is the value-iteration update equation 6 with dictionary size N_M and tables $Z^1, \dots, Z^{N_M}$; thus K_M iterations from zero approximate its restriction with error at most $\gamma^{K_M} V_{\max}$. For an arbitrary queried window, after ℓ transitions every table in the window belongs to the dictionary, so the backward recursion equation 48 and the score formula equation 49, run with dictionary $Z^1, \dots, Z^{N_M}$, return Q_M^{dict} exactly when initialized with the exact dictionary values (Lemma 7), and each backward step multiplies a terminal error by at most γ (Lemma 8). Consequently, uniformly over all queried windows and actions,

$$\|\widetilde{Q}_M - Q_M^{\mathrm{dict}}\|_\infty \le \gamma^{K_M+\ell} V_{\max} \le e^{-(1-\gamma)K_M} V_{\max} \le M^{-2},$$

by the choice $K_M = \lceil V_{\max} \log(V_{\max} M^2) \rceil$. Since $N_M \ge M V_{\max}^3$,

$$\gamma V_{\max}^2 \sqrt{\frac{\Lambda}{2N_M}} \le \sqrt{\frac{V_{\max}\Lambda}{2M}}.$$

The triangle inequality proves equation 85 at this update, conditionally on the observed tables. Removing the conditioning and taking a union bound over at most J updates gives the stated probability.

Proof of equation 86. Fix x and let $a^\star$ maximize $\widehat{Q}^\star_M(x, \cdot)$. Since $\pi_M(x)$ maximizes $\widetilde{Q}_M(x, \cdot)$,

$$0 \le \widehat{V}^\star_M(x) - \widehat{Q}^\star_M(x, \pi_M(x)) \le \big[\widehat{Q}^\star_M(x, a^\star) - \widetilde{Q}_M(x, a^\star)\big] + \big[\widetilde{Q}_M(x, \pi_M(x)) - \widehat{Q}^\star_M(x, \pi_M(x))\big]$$
$$\le 2\|\widetilde{Q}_M - \widehat{Q}^\star_M\|_\infty,$$

which is equation 68, and by equation 85 this is at most $2\big[\sqrt{V_{\max}\Lambda/(2M)} + M^{-2}\big]$. This is the empirical Bellman residual of π_M at x; it holds at every augmented state, so equation 82 with this residual b gives equation 86. □

The Bellman equation of the active policy decomposes each summand $(1-\gamma)W_t(X_t) - r_t$ into a difference of successive values and a martingale difference. Since W_t is fixed within each epoch, the value differences telescope; the following lemma controls the epoch boundaries and the martingale fluctuations.

Lemma 14 (Epochwise value-to-reward comparison)**.** *Suppose $\{0, \dots, T-1\}$ is partitioned into at most J deterministic consecutive epochs. At the start of each epoch, a deterministic stationary policy is selected using the information then available (possibly depending on previous observations and on the algorithm's randomization) and executed throughout the epoch. Let W_t be the true infinite-horizon value of the policy active at time t. For every $w > 0$, with probability at least $1 - 2e^{-w}$,*

$$\sum_{t=0}^{T-1} \big[(1-\gamma)W_t(X_t) - r_t\big] \le V_{\max} J + \sqrt{2\big(4V_{\max}T + 2V_{\max}^2 J + 16V_{\max}^2 w\big)w} + \tfrac{2}{3}V_{\max} w. \quad (88)$$

Proof. Let $\mathcal{F}_t$ be the σ-field of the information available after choosing A_t and before revealing the new table $\Theta_{t+\ell}$; in particular W_t and X_t are $\mathcal{F}_t$-measurable and $S_{t+1} = \Theta_t(S_t, A_t)$ is already determined. Since the active policy is executed at time t, its Bellman equation gives $W_t(X_t) = r_t + \gamma\, \mathbb{E}[W_t(X_{t+1}) \mid \mathcal{F}_t]$, hence

$$(1-\gamma)W_t(X_t) - r_t = \gamma\big[W_t(X_{t+1}) - W_t(X_t)\big] + \gamma\big[\mathbb{E}[W_t(X_{t+1}) \mid \mathcal{F}_t] - W_t(X_{t+1})\big]. \quad (89)$$

The first bracket telescopes within each epoch, because W_t is the same function throughout the epoch; as $0 \le W_t \le V_{\max}$, its total contribution is at most $V_{\max} J$.

The second bracket defines martingale differences with conditional mean zero and absolute value at most $V_{\max}$. Their total conditional variance is

$$\mathcal{V}_T := \sum_{t=0}^{T-1} \gamma^2 \operatorname{Var}\big(W_t(X_{t+1}) \mid \mathcal{F}_t\big). \tag{90}$$

We first bound $\mathcal{V}_T$. Using the Bellman equation, $0 \le r_t \le 1$ and $0 \le W_t \le V_{\max}$,

$$\begin{aligned}\gamma^2 \operatorname{Var}\big(W_t(X_{t+1}) \mid \mathcal{F}_t\big) &= \gamma^2 \mathbb{E}[W_t(X_{t+1})^2 \mid \mathcal{F}_t] - \big(W_t(X_t) - r_t\big)^2 \\ &\le \gamma^2 \big[W_t(X_{t+1})^2 - W_t(X_t)^2\big] + 2V_{\max} \\ &\quad + \gamma^2 \big[\mathbb{E}[W_t(X_{t+1})^2 \mid \mathcal{F}_t] - W_t(X_{t+1})^2\big].\end{aligned}$$

The squared-value differences telescope within each epoch and contribute at most $V_{\max}^2 J$, so

$$\mathcal{V}_T \le 2V_{\max} T + V_{\max}^2 J + \sum_{t=0}^{T-1} \gamma^2 \big[\mathbb{E}[W_t(X_{t+1})^2 \mid \mathcal{F}_t] - W_t(X_{t+1})^2\big]. \tag{91}$$

The last sum is a martingale with increments bounded by $V_{\max}^2$; since $z \mapsto z^2$ is $2V_{\max}$-Lipschitz on $[0, V_{\max}]$, its total conditional variance is at most $4V_{\max}^2 \mathcal{V}_T$. Applying $e^z \le 1 + z + z^2$ for $|z| \le 1$ to the increments divided by $8V_{\max}^2$, followed by the exponential supermartingale inequality, gives

$$\mathbb{P}\left\{ \sum_{t=0}^{T-1} \gamma^2 \big[\mathbb{E}[W_t(X_{t+1})^2 \mid \mathcal{F}_t] - W_t(X_{t+1})^2\big] > \frac{\mathcal{V}_T}{2} + 8V_{\max}^2 w \right\} \le e^{-w}.$$

Combined with equation 91, this shows that, except with probability e^{-w},

$$\mathcal{V}_T \le 4V_{\max} T + 2V_{\max}^2 J + 16 V_{\max}^2 w. \tag{92}$$

Finally, Freedman's inequality for martingales with increments bounded by $V_{\max}$ shows that, on the event equation 92, the sum of the second brackets in equation 89 is at most $\sqrt{2(4V_{\max} T + 2V_{\max}^2 J + 16 V_{\max}^2 w) w} + \frac{2}{3} V_{\max} w$ except with probability e^{-w}. Adding the telescoping contribution proves equation 88. □

We sum the policy losses in equation 64. On $\mathcal{E}_{\mathrm{model}} \cap \mathcal{E}_{\mathrm{dict}}$, equation 86 and $(1-\gamma)V_{\max} = 1$ imply, whenever $M_t \ge M_0$,

$$(1-\gamma)\big[V_\ell^\star(X_t) - W_t(X_t)\big] \le (8+\sqrt{2}) \sqrt{\frac{V_{\max}\Lambda}{M_t}} + \frac{8 V_{\max}\Lambda}{3 M_t} + \frac{2}{M_t^2}. \tag{93}$$

At $t = 0$ and at times with $M_t < M_0$, the normalized policy loss is at most 1; since $M_t \ge t/2$, there are at most $2M_0$ such times. For $t \ge 1$, $M_t \ge t/2$ and an epoch starting at M contains at most M steps, so

$$\sum_{t=1}^{T-1} \frac{1}{\sqrt{M_t}} \le 2\sqrt{2T}, \qquad \sum_{t=1}^{T-1} \frac{1}{M_t} \le J, \qquad \sum_{t=1}^{T-1} \frac{1}{M_t^2} \le 2. \tag{94}$$

Summing equation 93 therefore gives

$$(1-\gamma) \sum_{t=0}^{T-1} \big[V_\ell^\star(X_t) - W_t(X_t)\big] \le 2M_0 + 4 + (16\sqrt{2}+4)\sqrt{V_{\max} T \Lambda} + \frac{8}{3} V_{\max} J \Lambda. \tag{95}$$

For the value-to-reward error, the epochs are deterministic and the policy is fixed within each of them, so Lemma 14 applies. Taking $w = \log(16/\delta)$ gives, with probability at least $1 - \delta/8$,

$$\sum_{t=0}^{T-1} \big[(1-\gamma) W_t(X_t) - r_t\big] \le V_{\max} J + \sqrt{2\Big(4V_{\max} T + 2V_{\max}^2 J + 16 V_{\max}^2 \log\tfrac{16}{\delta}\Big) \log\tfrac{16}{\delta}} + \frac{2}{3} V_{\max} \log\frac{16}{\delta}. \tag{96}$$

This bound holds under the original probability law, without conditioning on $\mathcal{E}_{\mathrm{model}}$ or $\mathcal{E}_{\mathrm{dict}}$.

Substituting equation 95 and equation 96 into equation 64, and using $\sqrt{2(a+b+c)w} \le \sqrt{2aw} + \sqrt{2bw} + \sqrt{2cw}$, we obtain on an event of probability at least

$$1 - \Big(\frac{3\delta}{16} + \frac{\delta}{8} + \frac{\delta}{8}\Big) = 1 - \frac{7\delta}{16} \ge 1 - \delta$$

the bound

$$\begin{aligned}\mathrm{Reg}(T) \le 2M_0 + 4 + (16\sqrt{2}+4)\sqrt{V_{\max} T \Lambda} + \frac{8}{3} V_{\max} J \Lambda + V_{\max} J \\ + \sqrt{8 V_{\max} T \log\tfrac{16}{\delta}} + 2V_{\max}\sqrt{J \log\tfrac{16}{\delta}} + \big(\sqrt{32} + \tfrac{2}{3}\big) V_{\max} \log\tfrac{16}{\delta}.\end{aligned}$$

Recalling $\Lambda = \log(16J/\delta) + \big(1 + (\ell-1)|\mathcal{S}||\mathcal{A}|\big)\log|\mathcal{S}|$, $M_0 = \lceil 16 V_{\max}^2 \Lambda_0 \rceil$ with $\Lambda_0 \le |\mathcal{S}|^2|\mathcal{A}| \log 2 + \log(16J/\delta)$, $J = O(\log(T+1))$, and $V_{\max} = (1-\gamma)^{-1}$, this gives, for fixed ℓ,

$$\mathrm{Reg}(T) \le \widetilde{O}_\ell\left(\sqrt{\frac{|\mathcal{S}||\mathcal{A}|\,T}{1-\gamma}} + \frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^2}\right),$$

which is equation 3. This proves the theorem. □

Remark 3 (Conversion to the undiscounted return deficit)**.** *Let $R_t := \sum_{k\ge 0}\gamma^k r_{t+k}$ denote the realised discounted return from time t, and $\mathcal{R}_\ell(T) := \sum_{t=0}^{T-1}[V_\ell^\star(X_t) - R_t]$. The identity $r_t = R_t - \gamma R_{t+1}$ gives $\mathrm{Reg}(T) = (1-\gamma)\mathcal{R}_\ell(T) + \gamma(R_T - R_0)$, and since $0 \le R_t \le V_{\max}$,*

$$\mathcal{R}_\ell(T) \le V_{\max}\,\mathrm{Reg}(T) + \gamma V_{\max}^2. \tag{97}$$

B Experimental Protocol

We evaluate RPTAS on the wind-farm storage-control benchmark of Lu et al. (2025). We use the wind, price, and mismatch arrays produced by their preprocessing, together with the same state and action discretisations, battery dynamics, and evaluation interval. Code is available here

Control objective. At each time-step, the wind farm announces how much electricity it will supply to the grid. If its realised production is lower than the announced amount, it pays a penalty for the resulting shortfall. The battery can store energy when production is abundant and release it to compensate for future shortages. Since its capacity is limited, the agent must decide when to store or use energy based on the predicted prices and wind mismatches. Its objective is to minimise the expected discounted sum of imbalance penalties.The planning problem is therefore to allocate the limited battery capacity over time so as to minimise the expected discounted imbalance cost.

State and action spaces. In this benchmark, the committed generation is fixed in advance and is not a decision variable of the battery controller. Consequently, the price p_t and mismatch are unaffected by the battery action. We therefore group them into the uncontrolled state $z_t = (p_t, D_t)$.

$$e_t = 10\,b_p(p_t) + b_D(D_t) \in \{0, \dots, 99\}.$$

The battery state of charge belongs to $\mathcal{X} = \{0, 0.5, \dots, 10\}$, and the action set is $\mathcal{A} = \{-2, -1.5, \dots, 2\}$ kWh. Consequently, the complete MDP contains $100 \times 21 = 2{,}100$ states and nine actions. Given a battery level x and nominal action a, the feasible action and the next battery state are

$$\widetilde{a}(x,a) = \mathrm{clip}(a, -x, 10-x), \qquad x^+ = x + \widetilde{a}(x,a). \tag{98}$$

Because the battery grid and action increments both have spacing 0.5 kWh, the next battery state always belongs to $\mathcal{X}$ and no additional projection error is introduced. The realised one-step cost is

$$c(p, D, x, a) = p \max\{0, -(D + \widetilde{a}(x,a))\}.$$

All planning recursions use discount factor $\gamma = 0.95$ and are implemented directly in cost-minimisation form. Performance is evaluated using the undiscounted cumulative realised cost in the original units of equation B.

Evaluation trajectory. As in the BOLA experiment, policies are evaluated for $T = 3{,}000$ time steps, using indices $t = 1952, \ldots, 4951$, and an initial battery level of $x_{1952} = 5$ kWh. All methods start from the same battery level and face the same price and wind sequence. However, their states of charge may evolve differently because they choose different charging and discharging actions.

B.1 FORECAST

At each time step u, the forecast is generated as

$$\widehat{p}_u = p_u(1 + \sigma\varepsilon_u^p), \qquad \widehat{D}_u = D_u(1 + \sigma\varepsilon_u^D), \qquad \varepsilon_u^p, \varepsilon_u^D \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \tag{99}$$

The case $\sigma = 0$ corresponds to perfect predictions. We consider $\sigma \in \{0, .05, .10, .20, .30, .40, .50, .60, .80, 1\}$.

Following the convention of Lu et al. (2025), figures report the relative noise level as $100\sigma\%$. For each seed, we first generate one sequence of noise variables $\{(\varepsilon_u^p, \varepsilon_u^D)\}_u$, with one pair for each date u. This same sequence is used by every method and every look-ahead depth. For a noise level σ, the forecast at date u is obtained by multiplying these fixed noise variables by σ as in equation 99. Hence, two methods evaluated with the same seed and σ receive exactly the same forecast for any given date. Differences in performance therefore cannot be attributed to different random forecast errors.

Information convention. Throughout the experiments, ℓ denotes the number of exogenous observations available before acting. At date t, every method observes (p_t, D_t) exactly and receives predictions for $t+1, \ldots, t+\ell$. In particular, $\ell = 0$ is the classical MDP without look-ahead. The cost of the current action is always evaluated using the exact current price and mismatch, $c_t(x, a) = p_t \max\{0, -(D_t + \widetilde{a}(x, a))\}$.

Posterior kernels. RPTAS and BOLA-BAYES convert each noisy forecast into a posterior transition kernel. To approximate the continuous distribution within each bin, we sample without replacement $4{,}000$ historical pairs (p_i, D_i) from the training interval, using seed 7. These pairs, referred to as atoms, retain both their exogenous bin e_i and their exact cost under each feasible battery action.

B.2 EXOGENOUS MODEL AND POSTERIOR KERNELS

The exogenous transition model is estimated on a historical interval that starts $1{,}000$ observations after the end of the evaluation trajectory, so the training and evaluation intervals are disjoint. Price and mismatch transitions are estimated separately. For $q \in \{p, D\}$, we use

$$\widehat{P}_q(j \mid i) = \frac{1 + \sum_{t \in \mathcal{I}_{\text{train}}} \mathbf{1}\{b_q(q_t) = i,\ b_q(q_{t+1}) = j\}}{10 + \sum_{t \in \mathcal{I}_{\text{train}}} \mathbf{1}\{b_q(q_t) = i\}}, \tag{100}$$

where the additive terms correspond to one pseudo-count for each possible successor bin. The joint transition kernel is then

$$\widehat{P}_{\text{exo}}\big((i', j') \mid (i, j)\big) = \widehat{P}_p(i' \mid i)\widehat{P}_D(j' \mid j). \tag{101}$$

To convert a noisy forecast into a posterior over the next exogenous state, we sample without replacement $4{,}000$ historical pairs (p_r, D_r) from the training interval. Each pair, called an atom, is associated with its discrete bin e_r and its exact cost under every battery action. Given a forecast $\widehat{z}_u = (\widehat{p}_u, \widehat{D}_u)$, the likelihood weight of atom r is

$$w_{u,r} = \frac{1}{s_r^p s_r^D} \exp\left(-\frac{(\widehat{p}_u - p_r)^2}{2(s_r^p)^2} - \frac{(\widehat{D}_u - D_r)^2}{2(s_r^D)^2}\right), \tag{102}$$

where $s_r^p = \max\{\sigma|p_r|, 10^{-3}\Delta_p\}$, $s_r^D = \max\{\sigma|D_r|, 10^{-3}\Delta_D\}$, and Δ_p, Δ_D are the corresponding bin widths. The likelihood of a bin e' is the average weight of its atoms:

$$L_u(e') = \frac{1}{n_{e'}} \sum_{r : e_r = e'} w_{u,r}, \tag{103}$$

where $n_{e'}$ is the number of atoms in that bin. We use the average rather than the sum because the empirical frequency of the bin is already accounted for by the prior kernel $\widehat{P}_{\mathrm{exo}}$. Combining this likelihood with the transition prior gives

$$\kappa_u(e, e') = \frac{\widehat{P}_{\mathrm{exo}}(e, e')L_u(e')}{\sum_j \widehat{P}_{\mathrm{exo}}(e, j)L_u(j)}. \tag{104}$$

Future costs are estimated by averaging the exact atom costs using the same weights. For bins containing no atom, we use the cost evaluated at the bin centre. When $\sigma = 0$, the future exogenous bin is observed exactly and κ_u reduces to a point mass on that bin.

RPTAS and BOLA-BAYES use the same estimated transition kernel, posterior kernels, and posterior costs. The value of σ is assumed known when computing these quantities. The certainty-equivalent MPC and BOLA controllers do not construct a posterior; they plan directly with the point forecasts $(\widehat{p}_u, \widehat{D}_u)$.

B.3 COMPARED METHODS

We compare four controllers:

RPTAS. For each pair (ℓ, σ), we draw a dictionary of $N = 8$ posterior kernels using the model in equation 104. For every current exogenous bin e, a dictionary row is generated by sampling a successor bin from $\widehat{P}_{\mathrm{exo}}(e, \cdot)$, sampling an atom in that bin, corrupting it according to equation 99, and computing the resulting posterior row. Rows are generated independently; at $\sigma = 0$, they are one-hot. We then run the cost-minimisation version of Algorithm 1 on the dictionary state space $\mathcal{S} \times [N]^\ell$. Value iteration is initialised at zero and stopped when the sup-norm difference between consecutive iterates is below 10^{-8}, or after 250 iterations. At $\ell = 0$, we run standard value iteration directly under $\widehat{P}_{\mathrm{exo}}$. At time t, the online recursion receives $\kappa_{t+1}, \ldots, \kappa_{t+\ell}$, the associated posterior costs, and the exact current cost. RPTAS selects one action, observes the next state, and repeats the computation using the shifted forecast window. The dictionary is sampled with seed 0 and held fixed across the ten evaluation seeds.

BOLA-BAYES. This controller implements the two-stage BOLA principle described by Lu et al. (2025). Its offline stage computes a state-only Bayesian continuation value $J_{\ell,\sigma}(e, x)$. For every (ℓ, σ), the Bellman operator is approximated using $N_2 = 8$ synthetic prediction windows sampled from the same forecast model as the RPTAS dictionary. Its fixed-point iteration uses seed 0, tolerance 10^{-4}, and at most 500 iterations. Online, BOLA-BAYES uses the posterior kernels and posterior costs to select an action sequence for dates $t, \ldots, t + \ell$. The unobserved transition following date $t + \ell$ is integrated using $\widehat{P}_{\mathrm{exo}}$, after which $J_{\ell,\sigma}$ supplies the terminal cost. The complete sequence of $\ell + 1$ actions is executed before a new forecast window is requested.

BOLA and MPC. The controller denoted BOLA is the zero-terminal, certainty-equivalent block controller used in the wind-farm implementation of Lu et al. (2025), aligned with our information convention and common forecast stream. It uses the exact current cost and the point forecasts for dates $t + 1, \ldots, t + \ell$, computes an $\ell + 1$-action plan, and executes the entire plan.

MPC uses exactly the same finite-horizon recursion and zero terminal value but executes only its first action. It then replans at the next date using the shifted forecast window. Consequently, the MPC/BOLA comparison isolates the effect of committing to the complete block.

Our notation differs from that of the BOLA implementation: its horizon K counts the current date as part of the action block, whereas ℓ counts only strictly future observations. The corresponding parameters are therefore related by $K = \ell + 1$.

Scope of the comparison. RPTAS and BOLA-BAYES receive the same Bayesian forecast information and are the main algorithmic comparison. However, they differ both in execution and in their representation of future value: RPTAS uses a value indexed by the dictionary window and replans at every step, whereas BOLA-BAYES uses the state-only value $J_{\ell,\sigma}$ and commits to a block. Their comparison should therefore not be interpreted as an ablation of block commitment alone. That effect is isolated by the MPC/BOLA pair.

B.4 METRIC AND EVALUATION

The main sweep uses

$$\ell \in \{0, 1, 2, 3\}, \qquad \sigma \in \{0, .05, .10, .20, .30, .40, .50, .60, .80, 1\},$$

and forecast seeds $100, \ldots, 109$. For a fixed seed, all methods use the same realised price and wind-mismatch sequence; initial battery level and battery constraints; exact current price and mismatch; underlying noisy forecast at every absolute future date.

The methods do not necessarily make the same number of planning calls: RPTAS and MPC replan after every transition, while the BOLA controllers request a new window only after completing their current block. This is the intended distinction between receding-horizon control and block commitment. Nevertheless, whenever two methods use a prediction for the same absolute date, that prediction is generated from the same noise variables.

Offline quantities are recomputed for every (ℓ, σ) but are held fixed across the ten forecast seeds. Thus, the uncertainty bands in the main experiment quantify sensitivity to forecast noise conditional on the sampled dictionary and Bayesian terminal-value approximation; they do not include dictionary-sampling variability.

For a controller A, its final cumulative realised cost is

$$C_A = \sum_{t=1952}^{4951} c(p_t, D_t, x_t^A, a_t^A). \tag{105}$$

Following Lu et al. (2025), we report the cost reduction relative to their no-prediction reference:

$$\mathrm{CR}_A = C_{\mathrm{NP}} - C_A. \tag{106}$$

Higher values indicate better performance, and $\mathrm{CR}_A < 0$ means that the controller performs worse than the no-prediction reference.

The reference C_{NP} is computed using the no-prediction policy from the BOLA notebook. That policy performs 100 value-iteration updates while keeping the price and mismatch indices fixed and propagating only the battery state. We retain it because equation 106 is the reporting convention of the benchmark. It is distinct from the correctly modelled classical MDP at $\ell = 0$, which propagates the exogenous state using $\widehat{P}_{\mathrm{exo}}$. We therefore display the two quantities as separate reference lines.

For each (A, ℓ, σ), we report the mean over the ten common forecast seeds. Shaded regions correspond to

$$\text{mean} \ \pm \ 2\frac{\widehat{\mathrm{sd}}(C_A)}{\sqrt{10}}.$$

Since C_{NP} is deterministic, the same bands apply to the reported cost reductions.

C EXTENSION TO NOISY TRANSITION LOOK-AHEAD

This section extends our results to locally corrupted transition look-ahead. We first specify the observation model and the correct comparator, then give the planning algorithm and its guarantee, and finally the learning algorithm and its regret analysis.

Conventions. A tilde marks objects of the noisy model: $\widetilde{\Theta}_t$, $\widetilde{L}$, $\widetilde{P}$, $\widetilde{\mathcal{T}}_\ell$, $\widetilde{V}_\ell^\star$, $\widetilde{Q}_\ell^\star$. As in Appendices A.2 and A.3, N is a dictionary size, M a number of observed transitions, $\widehat{\widetilde{L}}_N$ the empirical law of a dictionary of N noisy tables and $\widehat{\widetilde{L}}_M$ the product law of empirical rows built from M observed noisy tables; the two are distinct objects. Computed action scores are written $\widetilde{Q}_v$ (Appendix C.2) and $\widetilde{Q}_M$ (Appendix C.3), following the convention of Appendices A.2–A.3; the star distinguishes them from the optimal scores $\widetilde{Q}_\ell^\star$. We write $V_{\max} := (1-\gamma)^{-1}$, $|\Omega| = |\mathcal{S}|^{|\mathcal{S}||\mathcal{A}|}$, and $\widetilde{\theta}_{k:k'} := (\widetilde{\theta}_k, \ldots, \widetilde{\theta}_{k'})$.

C.1 NOISY TRANSITION LOOK-AHEAD

Recall that $\Omega = \mathcal{S}^{\mathcal{S}\times\mathcal{A}}$ and that the transition tables are i.i.d. with product law

$$L(\theta) = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} P\big(\theta(s,a) \mid s,a\big). \tag{107}$$

For every (s,a,s'), let $I(\cdot \mid s,a,s')$ be a distribution on $\mathcal{S}$. Conditional on $\Theta_t = \theta$, the noisy table $\widetilde{\Theta}_t$ is sampled according to

$$\mathbb{P}\big(\widetilde{\Theta}_t = \widetilde{\theta} \mid \Theta_t = \theta\big) = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} I\big(\widetilde{\theta}(s,a) \mid s,a,\theta(s,a)\big). \tag{108}$$

The pairs $(\Theta_t, \widetilde{\Theta}_t)$ are independent across time. Before choosing A_t, the learner observes

$$\widetilde{C}_t^{\ell} := (\widetilde{\Theta}_t, \dots, \widetilde{\Theta}_{t+\ell-1}), \tag{109}$$

whereas the true transition remains $S_{t+1} = \Theta_t(S_t, A_t)$. After acting, the agent observes only S_{t+1}; in particular, it never observes the complete table Θ_t.

Let $\widetilde{L}$ be the marginal law of a noisy table; it is again a product distribution on Ω. The true successor conditional on the noisy entry is governed by

$$\widetilde{P}(s' \mid s,a,z) := \mathbb{P}\big(\Theta_t(s,a) = s' \mid \widetilde{\Theta}_t(s,a) = z\big). \tag{110}$$

Whenever the conditioning event has positive probability, Bayes' rule gives

$$\widetilde{P}(s' \mid s,a,z) = \frac{P(s' \mid s,a) I(z \mid s,a,s')}{\sum_{s''\in\mathcal{S}} P(s'' \mid s,a) I(z \mid s,a,s'')}, \tag{111}$$

and the posterior may be chosen arbitrarily on null conditioning events. The product assumptions imply the local sufficiency identity

$$\mathbb{P}\big(\Theta_t(s,a) = s' \mid \widetilde{\Theta}_t = \widetilde{\theta}\big) = \widetilde{P}\big(s' \mid s,a,\widetilde{\theta}(s,a)\big). \tag{112}$$

For $V : \mathcal{S} \times \Omega^{\ell} \to [0, V_{\max}]$, the Bellman optimality operator of the noisy augmented MDP is

$$(\widetilde{\mathcal{T}}_{\ell} V)(s, \widetilde{\theta}_{0:\ell-1}) := \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \widetilde{P}\big(s' \mid s,a,\widetilde{\theta}_0(s,a)\big) \mathbb{E}_{\widetilde{\Theta}\sim\widetilde{L}}\big[V(s', \widetilde{\theta}_{1:\ell-1}, \widetilde{\Theta})\big] \right\}. \tag{113}$$

It is a monotone γ-contraction mapping $[0, V_{\max}]$-valued functions to $[0, V_{\max}]$-valued functions. Its fixed point and associated action-value function are denoted $\widetilde{V}_{\ell}^{\star}$ and $\widetilde{Q}_{\ell}^{\star}$.

C.2 NEAR-OPTIMAL PLANNING

Suppose first that one can sample from $\widetilde{L}$ and evaluate $\widetilde{P}$. Draw

$$\widetilde{\Theta}^1, \dots, \widetilde{\Theta}^N \overset{\text{i.i.d.}}{\sim} \widetilde{L}, \qquad \widehat{\widetilde{L}}_N := \frac{1}{N} \sum_{j=1}^{N} \delta_{\widetilde{\Theta}^j}. \tag{114}$$

On $\mathcal{D}_N = \mathcal{S} \times [N]^{\ell}$, define the proxy operator, acting on arrays $v : \mathcal{D}_N \to \mathbb{R}$,

$$(\widetilde{\mathcal{T}}_N v)(s, i_0, \dots, i_{\ell-1}) := \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \widetilde{P}\big(s' \mid s,a,\widetilde{\Theta}^{i_0}(s,a)\big) \frac{1}{N} \sum_{j=1}^{N} v(s', i_1, \dots, i_{\ell-1}, j) \right\}, \tag{115}$$

let $\widetilde{V}_N^{\star}$ be its fixed point, and run $\widetilde{V}^0 := 0$, $\widetilde{V}^{k+1} := \widetilde{\mathcal{T}}_N \widetilde{V}^k$ for $k = 0, \dots, K-1$.

For an observed window $\widetilde{c} = (\widetilde{\theta}_0, \dots, \widetilde{\theta}_{\ell-1}) \in \Omega^{\ell}$ and an array $v : \mathcal{D}_N \to \mathbb{R}$, define $\widetilde{U}_{m,\widetilde{c}}^{v} : \mathcal{S} \times [N]^m \to \mathbb{R}$ by

$$\widetilde{U}_{\ell,\widetilde{c}}^{v}(u, i_0, \dots, i_{\ell-1}) := v(u, i_0, \dots, i_{\ell-1}), \tag{116}$$

and, for $m = \ell - 1, \dots, 1$,

$$\widetilde{U}^{v}_{m,\widetilde{c}}(u, i_0, \dots, i_{m-1}) := \max_{a' \in \mathcal{A}} \left\{ r(u,a') + \gamma \sum_{s' \in \mathcal{S}} \widetilde{P}\big(s' \mid u, a', \widetilde{\theta}_m(u,a')\big) \frac{1}{N} \sum_{i_m=1}^{N} \widetilde{U}^{v}_{m+1,\widetilde{c}}(s', i_0, \dots, i_m) \right\}. \tag{117}$$

The action scores at the root are

$$\widetilde{Q}_v(s, \widetilde{c}, a) := r(s,a) + \gamma \sum_{s' \in \mathcal{S}} \widetilde{P}\big(s' \mid s, a, \widetilde{\theta}_0(s,a)\big) \frac{1}{N} \sum_{i_0=1}^{N} \widetilde{U}^{v}_{1,\widetilde{c}}(s', i_0), \qquad a \in \mathcal{A}, \tag{118}$$

and the returned policy π_D is greedy with respect to $\widetilde{Q}_{\widetilde{V}^K}$. Relative to Algorithms 1–2, the deterministic successor $\theta(s,a)$ is replaced everywhere by the posterior average $\sum_{s'} \widetilde{P}(s' \mid s, a, \widetilde{\theta}(s,a))(\cdot)$.

Theorem 4 (RPTAS for noisy transition look-ahead)**.** *Fix $\ell \geq 2$ and $\varepsilon, \delta \in (0,1)$, and let N and K be as in equation 55. The algorithm above computes a policy π_D such that*

$$\mathbb{P}\Big(\widetilde{V}^{\pi_D}_{\ell}(s, \widetilde{c}) \geq \widetilde{V}^{\star}_{\ell}(s, \widetilde{c}) - \varepsilon V_{\max}, \quad \forall (s, \widetilde{c}) \in \mathcal{S} \times \Omega^{\ell} \Big) \geq 1 - \delta. \tag{119}$$

For every fixed ℓ, computing and storing the policy, as well as selecting an action at each decision time, require time and memory polynomial in $|\mathcal{S}|$, $|\mathcal{A}|$, ε^{-1}, $\log(1/\delta)$, and $(1-\gamma)^{-1}$.

Proof. The proof follows that of Theorem 2 step by step; we indicate the changes.

Ideal empirical model. Let $\widetilde{\mathcal{T}}_{\ell,N}$ be the operator equation 113 with $\mathbb{E}_{\widetilde{\Theta} \sim \widetilde{L}}$ replaced by $\mathbb{E}_{\widetilde{\Theta} \sim \widehat{\widetilde{L}}_N}$, and let $\widetilde{V}^{\star}_{\ell,N}$ and $\widetilde{Q}^{\star}_{\ell,N}$ be its fixed point and action-value function. Consider the class $\widetilde{\mathcal{F}} := \{\widetilde{\Theta} \mapsto \widetilde{V}^{\star}_{\ell}(u, \widetilde{\theta}_{1:\ell-1}, \widetilde{\Theta}) : u \in \mathcal{S},\ \widetilde{\theta}_{1:\ell-1} \in \Omega^{\ell-1}\}$, of cardinality at most $|\mathcal{S}||\Omega|^{\ell-1}$, whose members are deterministic and $[0, V_{\max}]$-valued. With $\eta := \varepsilon(1-\gamma)^2 V_{\max}/4$ as in equation 56, Lemma 3 applies verbatim: for the choice of N in equation 55, with probability at least $1-\delta$ every empirical average over the dictionary of a function of $\widetilde{\mathcal{F}}$ is within η of its expectation under $\widetilde{L}$. On this event, for every $(s, \widetilde{\theta}_{0:\ell-1}, a)$, the bracketed terms of $\widetilde{\mathcal{T}}_{\ell,N}\widetilde{V}^{\star}_{\ell}$ and $\widetilde{\mathcal{T}}_{\ell}\widetilde{V}^{\star}_{\ell}$ differ by γ times a $\widetilde{P}(\cdot \mid s, a, \widetilde{\theta}_0(s,a))$-average of such empirical-versus-true differences; averaging with respect to a probability distribution does not enlarge the error, so $\|\widetilde{\mathcal{T}}_{\ell,N}\widetilde{V}^{\star}_{\ell} - \widetilde{\mathcal{T}}_{\ell}\widetilde{V}^{\star}_{\ell}\|_{\infty} \leq \gamma\eta$. The arguments of Lemmas 4 and 5 then give

$$\|\widetilde{V}^{\star}_{\ell,N} - \widetilde{V}^{\star}_{\ell}\|_{\infty} \leq \frac{\gamma\eta}{1-\gamma}, \qquad \|\widetilde{Q}^{\star}_{\ell,N} - \widetilde{Q}^{\star}_{\ell}\|_{\infty} \leq \frac{\gamma\eta}{1-\gamma} \leq \frac{\varepsilon(1-\gamma)V_{\max}}{4}. \tag{120}$$

Restriction and extension. With the embedding ι of equation 41 built from $\widetilde{\Theta}^1, \dots, \widetilde{\Theta}^N$ and $R_N V := V \circ \iota$, the identity $R_N \widetilde{\mathcal{T}}_{\ell,N} V = \widetilde{\mathcal{T}}_N R_N V$ holds for every V, by the computation in the proof of Lemma 6 with the deterministic successor replaced by the posterior average; hence $\widetilde{V}^{\star}_{N} = R_N \widetilde{V}^{\star}_{\ell,N}$. The backward induction of Lemma 7, with the same replacement, shows that equation 117–equation 118 initialised with $v = \widetilde{V}^{\star}_{N}$ return $\widetilde{Q}_{\widetilde{V}^{\star}_{N}} = \widetilde{Q}^{\star}_{\ell,N}$ on every window $\widetilde{c} \in \Omega^{\ell}$. Finally, Lemma 8 holds unchanged because posterior averaging is a convex combination, hence 1-Lipschitz. Consequently, with $\|\widetilde{V}^K - \widetilde{V}^{\star}_{N}\|_{\infty} \leq \gamma^K V_{\max} \leq \varepsilon(1-\gamma)V_{\max}/4$ by the choice of K,

$$\sup_{s, \widetilde{c}, a} \big| \widetilde{Q}_{\widetilde{V}^K}(s, \widetilde{c}, a) - \widetilde{Q}^{\star}_{\ell}(s, \widetilde{c}, a) \big| \leq \frac{\gamma\eta}{1-\gamma} + \gamma^{\ell} \|\widetilde{V}^K - \widetilde{V}^{\star}_{N}\|_{\infty} \leq \frac{\varepsilon(1-\gamma)V_{\max}}{2}. \tag{121}$$

The approximate-greedy argument of the proof of Theorem 2, applied to $\widetilde{\mathcal{T}}_{\ell}$, proves equation 119.

Cost. Once the averages $\frac{1}{N}\sum_j v(s', i_1, \dots, i_{\ell-1}, j)$ are precomputed as in equation 45, each maximisation in equation 115 sums over $|\mathcal{S}|$ successors, so one sweep costs $O(|\mathcal{S}|(|\mathcal{S}||\mathcal{A}|+1)N^{\ell})$ and the backward extension $O(\ell|\mathcal{S}|(|\mathcal{S}||\mathcal{A}|+1)N^{\ell})$: relative to the perfect planner, the posterior sum introduces one additional factor $|\mathcal{S}|$. Memory is unchanged. □

C.3 Learning with noisy transition look-ahead

We use the observation model of Appendix C.1, with known rewards $r(s,a) \in [0,1]$. Neither P nor the corruption channel I needs to be known. Write $X_t = (S_t, \widetilde{C}_t^{\ell})$ and $r_t = r(S_t, A_t)$. The comparator observes the same noisy windows as the learner, and performance is measured by

$$\widetilde{\mathrm{Reg}}(T) := \sum_{t=0}^{T-1} \big[(1-\gamma)\widetilde{V}_{\ell}^{\star}(X_t) - r_t\big]. \tag{122}$$

Unlike perfect look-ahead, a noisy table does not reveal the physical successors, so the posterior $\widetilde{P}$ must be estimated from visited contexts and the planner must be optimistic. Fix a horizon $T \geq 2$ and a confidence level $\delta \in (0,1)$, and abbreviate

$$\Lambda_{\mathrm{P}} := \log \frac{8|\mathcal{S}|^2|\mathcal{A}|T}{\delta}, \qquad \Lambda_{\mathrm{D}} := \log \frac{8T|\mathcal{S}||\Omega|^{\ell-1}}{\delta}. \tag{123}$$

Estimators. At an update based on the first $M \geq 1$ observed transitions and noisy tables, define for every context $(s,a,z) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ the visit count and the empirical posterior

$$n_M(s,a,z) := \sum_{t=0}^{M-1} \mathbf{1}\{S_t = s,\ A_t = a,\ \widetilde{\Theta}_t(s,a) = z\}, \tag{124}$$

$$\widehat{\widetilde{P}}_M(s' \mid s,a,z) := \frac{\sum_{t=0}^{M-1} \mathbf{1}\{S_t = s,\ A_t = a,\ \widetilde{\Theta}_t(s,a) = z,\ S_{t+1} = s'\}}{n_M(s,a,z)}, \tag{125}$$

with an arbitrary probability vector when $n_M(s,a,z) = 0$. Since the noisy tables are observed in full, set

$$\widehat{\widetilde{L}}_M := \bigotimes_{(s,a)} \Big(\frac{1}{M} \sum_{t=0}^{M-1} \delta_{\widetilde{\Theta}_t(s,a)} \Big), \tag{126}$$

and define the exploration bonus

$$b_M(s,a,z) := \gamma V_{\max} \sqrt{\Lambda_{\mathrm{P}}} \left(\sqrt{\frac{|\mathcal{S}|}{1 \vee n_M(s,a,z)}} + \sqrt{\frac{|\mathcal{S}|^2|\mathcal{A}|}{M}} \right). \tag{127}$$

Optimistic dictionary planner. At each update, independently of previous dictionary draws conditional on the observations, sample

$$Z^1, \dots, Z^{N_M} \overset{\text{i.i.d.}}{\sim} \widehat{\widetilde{L}}_M, \qquad N_M := \big\lceil 2V_{\max}^4 M^2 \Lambda_{\mathrm{D}} \big\rceil, \qquad K_M := \lceil V_{\max} \log(2V_{\max} M) \rceil. \tag{128}$$

Starting from $v^0 := 0$ on $\mathcal{S} \times [N_M]^{\ell}$, perform K_M iterations of the clipped optimistic update

$$v^{k+1}(s, i_0, \dots, i_{\ell-1}) := \min\Bigg\{ V_{\max},\ \max_{a \in \mathcal{A}} \Bigg[r(s,a) + b_M\big(s,a,Z^{i_0}(s,a)\big) + \gamma \sum_{s' \in \mathcal{S}} \widehat{\widetilde{P}}_M\big(s' \mid s,a,Z^{i_0}(s,a)\big) \frac{1}{N_M} \sum_{j=1}^{N_M} v^k(s', i_1, \dots, i_{\ell-1}, j) \Bigg] \Bigg\}. \tag{129}$$

For an observed window $\widetilde{c}$, run the backward extension equation 116–equation 118 with dictionary $Z^1, \dots, Z^{N_M}$, terminal array v^{K_M}, posterior $\widehat{\widetilde{P}}_M$, stage reward $r(u,a') + b_M(u,a',\widetilde{\theta}_m(u,a'))$ at level m, and clipping at $V_{\max}$ after each maximisation. Denote by $\widetilde{Q}_M(x,a)$ the resulting root score, i.e. the expression inside the root maximisation, before clipping, with the root action fixed. The deployed policy π_M selects an action maximising $\widetilde{Q}_M(x,\cdot)$ with a fixed tie-breaking rule.

Schedule. Choose any action at time 0. Before choosing the action at time $t \geq 1$, recompute the planner with $M = t$ if t is a power of two, or if some count $n_t(s, a, z)$ has just reached a power of two (including 1); otherwise keep the previous planner. Counts are updated after every observed transition even when the planner is not recomputed. For $t \geq 1$, let M_t denote the most recent update time; the policy π_{M_t} is fixed between updates.

Theorem 5 (Regret with noisy look-ahead)**.** *Under the local corruption model, for every fixed $\ell \geq 1$, $T \geq 2$, and $\delta \in (0, 1)$, the preceding algorithm satisfies, with probability at least $1 - \delta$,*

$$\widetilde{\mathrm{Reg}}(T) \leq \widetilde{O}_\ell \left(\frac{\sqrt{|\mathcal{S}|^3 |\mathcal{A}|\, T}}{1-\gamma} + \frac{|\mathcal{S}|^2 |\mathcal{A}|}{1-\gamma} \right), \tag{130}$$

where $\widetilde{O}_\ell$ hides logarithmic factors in $|\mathcal{S}|, |\mathcal{A}|, T, \delta^{-1}$ and $(1-\gamma)^{-1}$, and factors depending on ℓ. For fixed ℓ, the update time, memory, and per-decision computation are polynomial in $|\mathcal{S}|, |\mathcal{A}|, T, (1-\gamma)^{-1}$ and $\log(1/\delta)$.

Proof. 1. Uniform model confidence. We claim that, with probability at least $1-\delta/2$, simultaneously for all $1 \leq M \leq T$ and all (s, a, z),

$$\|\widehat{\widetilde{P}}_M(\cdot \mid s, a, z) - \widetilde{P}(\cdot \mid s, a, z)\|_1 \leq 2\sqrt{\frac{|\mathcal{S}|\, \Lambda_{\mathrm{P}}}{1 \vee n_M(s, a, z)}}, \qquad \|\widehat{\widetilde{L}}_M - \widetilde{L}\|_1 \leq 2\sqrt{\frac{|\mathcal{S}|^2 |\mathcal{A}|\, \Lambda_{\mathrm{P}}}{M}}. \tag{131}$$

For the first inequality, condition on the entire noisy-table sequence. By local sufficiency equation 112 and independence across dates and entries, each visit to a context (s, a, z) reveals a fresh successor with law $\widetilde{P}(\cdot \mid s, a, z)$, independent of previously revealed successors. Equivalently, successive visits to a context consume successive elements of an i.i.d. stream attached to that context, and this representation remains valid for adaptive actions because the action is chosen before its latent successor is revealed. For a fixed stream and a fixed prefix length $q \geq 1$, let $\widehat{p}_q$ be the empirical distribution of the first q elements and $p = \widetilde{P}(\cdot \mid s, a, z)$. Since $\|\widehat{p}_q - p\|_1 = \max_{\varsigma \in \{-1,1\}^{\mathcal{S}}} \varsigma^\top (\widehat{p}_q - p)$ and each $\varsigma^\top (\widehat{p}_q - p)$ is an average of q independent centred variables in $[-1, 1]$, Hoeffding's inequality and a union bound over the $2^{|\mathcal{S}|}$ sign vectors give, for every $\zeta > 0$,

$$\mathbb{P}\{\|\widehat{p}_q - p\|_1 > \zeta\} \leq 2^{|\mathcal{S}|} \exp(-q\zeta^2/2).$$

Taking $\zeta = 2\sqrt{|\mathcal{S}|\Lambda_{\mathrm{P}}/q}$ and a union bound over the $|\mathcal{S}|^2|\mathcal{A}|$ streams and $q \leq T$ gives failure probability at most $\delta/4$; this covers the random prefix lengths $n_M(s, a, z)$ without conditioning on their values. The case $n_M(s, a, z) = 0$ follows from $\|\widehat{p}_0 - p\|_1 \leq 2 \leq 2\sqrt{|\mathcal{S}|\Lambda_{\mathrm{P}}}$.

For the second inequality, the first M noisy tables are i.i.d. with product law $\widetilde{L}$. For a single row, the method of types gives $\mathbb{E} \exp[M\, \mathrm{KL}(\widehat{p}_M \| p)] \leq (M+1)^{|\mathcal{S}|}$, since the number of types of M samples on $|\mathcal{S}|$ categories is at most $(M+1)^{|\mathcal{S}|}$ and each type has probability at most $e^{-M\,\mathrm{KL}}$. Rows are independent and KL is additive over products, so $\mathbb{E} \exp[M\, \mathrm{KL}(\widehat{\widetilde{L}}_M \| \widetilde{L})] \leq (M+1)^{|\mathcal{S}|^2|\mathcal{A}|}$, and Markov's inequality yields, for every $w > 0$,

$$\mathbb{P}\big\{M\, \mathrm{KL}(\widehat{\widetilde{L}}_M \| \widetilde{L}) > w\big\} \leq (M+1)^{|\mathcal{S}|^2|\mathcal{A}|} e^{-w}. \tag{132}$$

Take $w = 2|\mathcal{S}|^2|\mathcal{A}|\Lambda_{\mathrm{P}}$. Since $\log(M+1) \leq \Lambda_{\mathrm{P}}$, the right-hand side is at most $(\delta/(8|\mathcal{S}|^2|\mathcal{A}|T))^{|\mathcal{S}|^2|\mathcal{A}|} \leq \delta/(4T)$, and a union bound over $M \leq T$ gives failure probability at most $\delta/4$. Pinsker's inequality, $\|\widehat{\widetilde{L}}_M - \widetilde{L}\|_1 \leq \sqrt{2\,\mathrm{KL}} \leq \sqrt{2w/M}$, proves the second inequality of equation 131.

2. Optimism and finite-planner accuracy. For the analysis, let $\overline{\mathcal{T}}_M$ be the full-space version of equation 129, acting on $V : \mathcal{S} \times \Omega^\ell \to [0, V_{\max}]$ with the dictionary average replaced by $\mathbb{E}_{\widetilde{\Theta} \sim \widehat{\widetilde{L}}_M}$, let $\overline{V}_M$ be its fixed point and $\overline{Q}_M$ its unclipped action scores. This operator is monotone, a γ-contraction, and maps $[0, V_{\max}]$-valued functions into $[0, V_{\max}]$-valued functions. For any V with values in $[0, V_{\max}]$ and any context (s, a, z), adding and subtracting the continuation computed with

the true posterior and the empirical table law, and using $|(p-p')^\top f| \le \frac{1}{2}\|p-p'\|_1 \operatorname{span}(f)$, the discrepancy between the empirical and true discounted continuation values is at most

$$\frac{\gamma V_{\max}}{2}\Big(\|\widehat{\widetilde{P}}_M(\cdot \mid s,a,z) - \widetilde{P}(\cdot \mid s,a,z)\|_1 + \|\widehat{\widetilde{L}}_M - \widetilde{L}\|_1\Big) \le b_M(s,a,z)$$

on the event equation 131, uniformly in V, including data-dependent V. Consequently $\overline{\mathcal{T}}_M \widetilde{V}^\star_\ell \ge \widetilde{V}^\star_\ell$, and monotonicity together with contraction imply

$$0 \le \widetilde{V}^\star_\ell \le \overline{V}_M \le V_{\max}. \tag{133}$$

Conditionally on the history before an update, $\overline{V}_M$ is fixed and the dictionary samples $Z^1,\dots,Z^{N_M}$ are i.i.d. from $\widehat{\widetilde{L}}_M$. Hoeffding's inequality for each of the at most $|\mathcal{S}||\Omega|^{\ell-1}$ functions $Z \mapsto \overline{V}_M(s', \widetilde{\theta}_{1:\ell-1}, Z)$, and a union bound, give

$$\sup_{s'\in\mathcal{S},\, \widetilde{\theta}_{1:\ell-1}\in\Omega^{\ell-1}} \Big| \frac{1}{N_M}\sum_{j=1}^{N_M} \overline{V}_M(s', \widetilde{\theta}_{1:\ell-1}, Z^j) - \mathbb{E}_{\widetilde{\Theta}\sim\widehat{\widetilde{L}}_M}\big[\overline{V}_M(s', \widetilde{\theta}_{1:\ell-1}, \widetilde{\Theta})\big]\Big| \le \frac{1}{2V_{\max}M} \tag{134}$$

except with conditional probability at most $\delta/(4T)$, by the choice of N_M in equation 128. A union bound over the at most T possible update times makes equation 134 hold at every update with failure probability at most $\delta/4$.

Posterior averaging and clipping are 1-Lipschitz, so equation 134 bounds the difference between the full-space operator $\overline{\mathcal{T}}_M$ and the dictionary operator equation 129, both evaluated at $\overline{V}_M$, by $\gamma/(2V_{\max}M)$; contraction bounds the difference between their fixed points by $\gamma/(2M)$, and the unclipped action scores then differ by at most $\gamma\big(\gamma/(2M) + 1/(2V_{\max}M)\big) = \gamma/(2M)$. The array v^{K_M} approximates the restriction of the dictionary fixed point with error at most $\gamma^{K_M}V_{\max}$; after ℓ transitions every table of a queried window belongs to the dictionary, so the backward extension recovers the exact dictionary scores from the exact array (Lemma 7, with posterior averaging, bonus and clipping, all 1-Lipschitz), and with v^{K_M} the additional score error is at most $\gamma^{K_M+\ell}V_{\max} \le 1/(2M)$ by the choice of K_M. Thus $\|\widetilde{Q}_M - \overline{Q}_M\|_\infty \le 1/M$, and since π_M maximises $\widetilde{Q}_M$,

$$\overline{V}_M(x) \le \overline{Q}_M(x, \pi_M(x)) + \frac{2}{M} \qquad \text{for every augmented state } x. \tag{135}$$

This remains true when the maximal score exceeds $V_{\max}$, since $\overline{V}_M(x) = \min\{V_{\max}, \max_a \overline{Q}_M(x,a)\}$.

3. Regret decomposition. Let $\mathcal{F}_t$ contain the observations and algorithmic randomisation available after choosing A_t, before observing S_{t+1} and the new noisy table $\widetilde{\Theta}_{t+\ell}$. Write $z_t := \widetilde{\Theta}_t(S_t, A_t)$. By equation 135 and the uniform discrepancy bound of step 2 applied to $\overline{V}_{M_t}$,

$$\overline{V}_{M_t}(X_t) - r_t - \gamma\,\mathbb{E}[\overline{V}_{M_t}(X_{t+1}) \mid \mathcal{F}_t] \le 2b_{M_t}(S_t, A_t, z_t) + \frac{2}{M_t}.$$

Together with equation 133, this yields

$$(1-\gamma)\widetilde{V}^\star_\ell(X_t) - r_t \le 2b_{M_t}(S_t, A_t, z_t) + \frac{2}{M_t} + \gamma\big[\overline{V}_{M_t}(X_{t+1}) - \overline{V}_{M_t}(X_t)\big] \tag{136}$$

$$+ \gamma\big[\mathbb{E}[\overline{V}_{M_t}(X_{t+1}) \mid \mathcal{F}_t] - \overline{V}_{M_t}(X_{t+1})\big]. \tag{137}$$

There are at most $J := 1 + (|\mathcal{S}|^2|\mathcal{A}|+1)(1+\lfloor \log_2 T\rfloor)$ epochs, including time 0. The value differences telescope within each epoch and contribute at most $V_{\max}$ per epoch. The last bracket defines martingale differences bounded by $\gamma V_{\max}$ in absolute value; Azuma's inequality bounds their sum by $\gamma V_{\max}\sqrt{2T\log(4/\delta)}$ with failure probability at most $\delta/4$, under the original law and without conditioning on the confidence events.

4. Summation. The update schedule ensures $M_t \ge t/2$ and $1 \vee n_{M_t}(s,a,z) \ge \frac{1}{2}\big(1 \vee n_t(s,a,z)\big)$ for all (s,a,z). Grouping visits by their $|\mathcal{S}|^2|\mathcal{A}|$ possible contexts, using $\sum_{j=0}^{q-1}(1\vee j)^{-1/2} \le 2\sqrt{q}$ and Cauchy–Schwarz,

$$\sum_{t=1}^{T-1} \frac{1}{\sqrt{1 \vee n_{M_t}(S_t, A_t, z_t)}} \le 2\sqrt{2|\mathcal{S}|^2|\mathcal{A}|\,T}, \qquad \sum_{t=1}^{T-1}\frac{1}{\sqrt{M_t}} \le 2\sqrt{2T}, \qquad \sum_{t=1}^{T-1}\frac{1}{M_t} \le 2(1+\log T).$$

Consequently,

$$\sum_{t=1}^{T-1} b_{M_t}(S_t, A_t, z_t) \le 2\sqrt{2}\,\gamma V_{\max}\sqrt{\Lambda_{\mathrm{P}}}\Big(\sqrt{|\mathcal{S}|^3|\mathcal{A}|\,T} + \sqrt{|\mathcal{S}|^2|\mathcal{A}|\,T}\Big).$$

The regret at time 0 is at most 1. Summing equation 137 therefore gives, on the intersection of the confidence events,

$$\begin{aligned}\widetilde{\mathrm{Reg}}(T) \le 1 + 4\sqrt{2}\,\gamma V_{\max}\sqrt{\Lambda_{\mathrm{P}}}\Big(\sqrt{|\mathcal{S}|^3|\mathcal{A}|\,T} + \sqrt{|\mathcal{S}|^2|\mathcal{A}|\,T}\Big) + 4(1+\log T)\\ + V_{\max} J + \gamma V_{\max}\sqrt{2T\log(4/\delta)},\end{aligned}$$

which implies equation 130. The failure probabilities are at most $\delta/2$ for the model estimates, $\delta/4$ for the dictionary approximations, and $\delta/4$ for the martingale, for a total of at most δ.

Finally, $\log|\Omega| = |\mathcal{S}||\mathcal{A}|\log|\mathcal{S}|$, so Λ_{D}, N_M and K_M are polynomial in the stated parameters for fixed ℓ. Each dictionary has $|\mathcal{S}|N_M^{\ell}$ states, the posterior sums have $|\mathcal{S}|$ terms, and the backward extension has depth ℓ; the claimed computational bounds follow as in Theorem 4. □