Span-Agnostic Optimal Sample Complexity and Oracle Inequalities for Average-Reward RL

Matthew Zurek MATTHEW.ZUREK@WISC.EDU and **Yudong Chen** YUDONGCHEN@CS.WISC.EDU
Department of Computer Sciences, University of Wisconsin-Madison

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We study the sample complexity of finding an ε-optimal policy in average-reward Markov Decision Processes (MDPs) with a generative model. The minimax optimal span-based complexity of $\widetilde{O}(SAH/\varepsilon^2)$, where H is the span of the optimal bias function, has only been achievable with prior knowledge of the value of H. Prior-knowledge-free algorithms have been the objective of intensive research, but several natural approaches provably fail to achieve this goal. We resolve this problem, developing the first algorithms matching the optimal span-based complexity without H knowledge, both when the dataset size is fixed and when the suboptimality level ε is fixed. Our main technique combines the discounted reduction approach with a method for automatically tuning the effective horizon based on empirical confidence intervals or lower bounds on performance, which we term *horizon calibration*. We also develop an *empirical span penalization* approach, inspired by sample variance penalization, which satisfies an *oracle inequality* performance guarantee. In particular this algorithm can outperform the minimax complexity in benign settings such as when there exist near-optimal policies with span much smaller than H.

Keywords: Reinforcement learning, average-reward MDPs, sample complexity, adaptivity

1. Introduction

Reinforcement Learning (RL) has achieved significant empirical successes in various fields, demonstrating its potential to solve complex decision-making problems. RL is commonly modeled as to learn a policy which maximizes cumulative rewards within a Markov Decision Process (MDP), where the cumulative rewards can be measured in several different ways. We focus on the average-reward criterion, which involves the long-term average of collected rewards as the horizon goes to infinity, making it suitable for ongoing tasks without a natural endpoint.

A fundamental question in average-reward RL is the sample complexity for learning a near-optimal policy under a generative model of the MDP. This question has been the subject of intensive research. Recent work has established the minimax-optimal span-based complexity $\widetilde{O}(SA\|h^\star\|_{\text{span}}/\varepsilon^2)$ for learning an ε-optimal policy (Zurek and Chen, 2025), where $\|h^\star\|_{\text{span}}$ denotes the span of the optimal bias $h^\star$ and is known to be a more refined complexity parameter than alternatives such as diameter or mixing times. However, this algorithm as well as earlier work all require prior knowledge of $\|h^\star\|_{\text{span}}$ (or other complexity parameters), which is generally unavailable, making the algorithms impractical. A flurry of subsequent research (Neu and Okolo, 2024; Tuynman et al., 2024; Jin et al., 2024; Zurek and Chen, 2024) has focused on removing the need for prior knowledge but failed to match the optimal span-based complexity. In fact, several natural approaches to knowledge-free optimal complexity, including span estimation and the average-reward plug-in method, are shown to provably fail (Tuynman et al., 2024; Zurek and Chen, 2024).

In this paper we resolve this problem, providing algorithms which obtain the optimal span-based complexity without knowing $\|h^\star\|_{\text{span}}$, for both settings where we fix the dataset size n and where

we prescribe a target suboptimality level ε. Our algorithms are based upon reductions to discounted MDPs (DMDPs) combined with a novel technique of (effective-)*horizon calibration*, which chooses discount factors to maximize lower bounds or minimize confidence intervals on policy performance. This technique can be seen as related, but representing a simpler alternative, to the technique of sample variance penalization (SVP) from statistical learning (Maurer and Pontil, 2009; Duchi and Namkoong, 2019). We further develop an algorithm based more closely on a relaxed version of SVP, which we call *empirical span penalization*, which enjoys even stronger guarantees. In particular, this algorithm satisfies a complexity bound in terms of the minimum span $\|h^\pi\|_{\text{span}}$ of any gain-optimal policy π in place of $\|h^\star\|_{\text{span}}$. Moreover, it adapts to and competes with simpler (potentially suboptimal) policies with the best tradeoff between complexity and suboptimality. This bound is reminiscent of the *oracle inequalities* from the statistical learning literature (Deheuvels et al., 2007; Koltchinskii, 2011), but to the best of our knowledge is new to average-reward RL.

1.1. Related Work

The problem of learning optimal policies in average-reward MDPs (AMDPs) is studied in Jin and Sidford (2020, 2021); Li et al. (2024); Wang et al. (2022); Zhang and Xie (2023); Wang et al. (2023). We start with the recent work Zurek and Chen (2025), which was the first to obtain the optimal span-based sample complexity but required prior knowledge of $\|h^\star\|_{\text{span}}$ to do so. All earlier work also required knowledge of problem-dependent complexity parameters. More recent work, which we discuss below, has studied the setting without such prior knowledge; see Table 1 for a summary.

Tuynman et al. (2024) and Zurek and Chen (2025) show that it is generally impossible to obtain a multiplicative estimate of $\|h^\star\|_{\text{span}}$ with $\text{poly}(SA\|h^\star\|_{\text{span}})$ samples. See Appendix B for discussion of the relationship between our algorithms and estimating $\|h^\star\|_{\text{span}}$. By estimating the MDP's diameter D, which upper bounds $\|h^\star\|_{\text{span}}$ but can be arbitrarily larger (Bartlett and Tewari, 2009; Lattimore and Szepesvári, 2020), the work in Tuynman et al. (2024) removes the need for prior knowledge within the algorithm of Zurek and Chen (2025) but obtains a complexity involving D rather than $\|h^\star\|_{\text{span}}$. The Q-learning-based algorithm in Jin et al. (2024) uses increasing discount factors and does not require prior knowledge. Their complexity bound however depends on the largest mixing time of all policies, τ_{unif}, which satisfies $3\tau_{\text{unif}} \geq \|h^\star\|_{\text{span}}$ and can be infinite or arbitrarily larger than $\|h^\star\|_{\text{span}}$ (Wang et al., 2022; Zurek and Chen, 2024). (See Appendix A.2 for definitions of D and τ_{unif}.)

Neu and Okolo (2024) and Zurek and Chen (2024) study, respectively, approaches based on stochastic saddle-point optimization and the average-reward plug-in method, both obtaining bounds involving the bias spans of certain policies output by the algorithm. These spans are not generally controlled by $\|h^\star\|_{\text{span}}$; in particular, Zurek and Chen (2024, Theorem 14) present an example where this is the case and show that the average-reward plug-in approach cannot achieve the optimal $SA\|h^\star\|_{\text{span}}/\varepsilon^2$ complexity. Zurek and Chen (2024) also analyzes a DMDP-reduction algorithm that uses a (relatively small) effective horizon independent of $\|h^\star\|_{\text{span}}$, achieving a suboptimal complexity with $\|h^\star\|_{\text{span}}^2$ dependence.

While for consistency we present all results in Table 1 in terms of the sample complexity sufficient for ε suboptimality, we note that only the results of Zurek and Chen (2025), Tuynman et al. (2024), and our Theorem 2 are for the fixed-ε setting (where a target suboptimality ε is provided to the algorithm); the others results are for the fixed-n setting (where the dataset size is n). We provide more discussion of the relationships between these two settings in Appendix A.3.

Also related to the present work are papers studying the sample complexity of the model-based/plug-in approach for discounted MDPs (Azar et al., 2012, 2013; Agarwal et al., 2020; Li et al., 2020; Zurek and Chen, 2024). We also note that Boone and Zhang (2024) recently developed an algorithm for the *online* setting achieving a $\|h^\star\|_{\text{span}}$-based regret bound without requiring prior knowledge. This result does not imply any sample complexity bounds in our setting, because there is no general regret-to-PAC conversion for average-reward MDPs (Tuynman et al., 2024), and even if this were possible, their result appears to require $\Omega(S^{40}A^{20}\|h^\star\|_{\text{span}}^{10})$ interaction steps before achieving the optimal regret, which would imply a massive "burn-in" cost in our setting.

2. Problem Setup

A Markov decision process is a tuple $(\mathcal{S}, \mathcal{A}, P, r)$, where $\mathcal{S}$, $\mathcal{A}$ are the state and action spaces, respectively, with finite cardinalities $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel with $\Delta(\mathcal{S})$ denoting the probability simplex on $\mathcal{S}$, and $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function. We only consider Markovian stationary policies of the form $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. For initial state $s_0 \in \mathcal{S}$ and policy π, let $\mathbb{E}_{s_0}^\pi$ denote the expectation w.r.t. the distribution over trajectories $(S_0, A_0, S_1, A_1, \dots)$ with $S_0 = s_0$, $A_t \sim \pi(S_t)$, and $S_{t+1} \sim P(\cdot \mid S_t, A_t)$. Let P_π denote the transition probability matrix of the Markov chain induced by π, where $(P_\pi)_{s,s'} := \sum_{a\in\mathcal{A}} \pi(a|s)P(s' \mid s, a)$. Similarly let $(r_\pi)_s := \sum_{a\in\mathcal{A}} \pi(a|s)r(s, a)$. We also consider P as an $(S \times A)$-by-S matrix with $P_{sa,s'} = P(s' \mid s, a)$, and r as an SA-dimensional vector. For a policy π, define the policy matrix $M^\pi \in \mathbb{R}^{S\times SA}$ by $M^\pi_{s,sa} = \pi(a|s)$ and $M^\pi_{s,s'a} = 0$ if $s \neq s'$. Note that $P_\pi = M^\pi P$ and $r_\pi = M^\pi r$. Also define the maximization operator $M : \mathbb{R}^{SA} \to \mathbb{R}^S$ by $M(x)_s = \max_a x_{sa}$.

We assume P is unknown, but one has access to a generative model (a.k.a. simulator) (Kearns and Singh, 1998), which provides independent samples from $P(\cdot \mid s, a)$ for each $s \in \mathcal{S}, a \in \mathcal{A}$. We assume r is known, which is standard (Agarwal et al., 2020; Li et al., 2020) as otherwise estimating r is relatively easy. Let $\mathbf{0}, \mathbf{1} \in \mathbb{R}^{\mathcal{S}}$ be the all-zero and all-one vectors, respectively.

Discounted reward criterion A discounted MDP is a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\gamma \in (0, 1)$ is the discount factor. For a policy π, the (discounted) value function $V_\gamma^\pi : \mathcal{S} \to [0, \infty)$ is defined as $V_\gamma^\pi(s) := \mathbb{E}_s^\pi\left[\sum_{t=0}^\infty \gamma^t R_t\right]$, where $R_t = r(S_t, A_t)$ is the reward received at time t. There always exists an optimal policy $\pi_\gamma^\star$ that satisfies $V_\gamma^{\pi_\gamma^\star}(s) = V_\gamma^\star(s) := \sup_\pi V_\gamma^\pi(s), \forall s \in \mathcal{S}$ (Puterman, 1994). When using transition kernel $\widehat{P}$ we will accordingly write $\widehat{V}_\gamma^\pi$ for the associated value function. For reward functions r' other than r, we include the reward function in the subscript e.g. $V_{\gamma,r'}^\pi$.

Average-reward criterion In an MDP $(\mathcal{S}, \mathcal{A}, P, r)$, the average reward, a.k.a. the *gain*, of a policy π starting from state s is defined as $\rho^\pi(s) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_s^\pi\left[\sum_{t=0}^{T-1} R_t\right]$. The *bias function* of a stationary policy π is $h^\pi(s) := \text{C-lim}_{T\to\infty} \mathbb{E}_s^\pi\left[\sum_{t=0}^{T-1} (R_t - \rho^\pi(S_t))\right]$, where C-lim denotes the Cesaro limit. When the Markov chain induced by P_π is aperiodic, C-lim can be replaced with the usual limit. A policy $\pi^\star$ is *Blackwell-optimal* if there exists some discount factor $\bar\gamma \in [0, 1)$ such that for all $\gamma \geq \bar\gamma$ we have $V_\gamma^{\pi^\star} \geq V_\gamma^\pi, \forall\pi$. When S and A are finite, there always exists some Blackwell-optimal policy, denoted by $\pi^\star$ (Puterman, 1994). Define the optimal gain $\rho^\star \in \mathbb{R}^{\mathcal{S}}$ by $\rho^\star(s) = \sup_\pi \rho^\pi(s)$ and note that $\rho^\star := \rho^{\pi^\star}$. Define the optimal bias $h^\star := h^{\pi^\star}$ (which is unique even when $\pi^\star$ is not). A policy π is *gain-optimal* if $\rho^\pi = \rho^\star$ and it is *bias-optimal* if in addition $h^\pi = h^\star$. For $x \in \mathbb{R}^{\mathcal{S}}$, define the span semi-norm $\|x\|_{\text{span}} := \max_{s\in\mathcal{S}} x(s) - \min_{s\in\mathcal{S}} x(s)$.

An MDP is communicating if for any states s and s', some policy can reach s' from s with probability 1. An MDP is weakly communicating if the states can be partitioned into two subsets $\mathcal{S} =$

Algorithm	Sample Complexity	Reference	Prior Knowledge
DMDP Reduction	$SA\frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2}$	Zurek and Chen (2025)	Yes
Diameter Estimation + DMDP Reduction	$SA\frac{D}{\varepsilon^2}+S^2AD^2$	Tuynman et al. (2024)	No
Dynamic Horizon Q-Learning	$SA\frac{\tau_{\text{unif}}^8}{\varepsilon^8}$	Jin et al. (2024)	No
Stochastic Saddle-Point Optimization	$S^2A^2\frac{\|h^{\widehat{\pi}}\|_{\text{span}}^4}{\varepsilon^2}$	Neu and Okolo (2024)	No
Plug-in Approach with Anchoring and Reward Perturbation	$SA\frac{\min\{D,\tau_{\text{unif}}\}}{\varepsilon^2}$ $SA\frac{\|h^\star\|_{\text{span}}+\min\left\{\left\|\widehat{\underline{h}}^\star\right\|_{\text{span}},\left\|\underline{h}^{\widehat{\pi}}\right\|_{\text{span}}\right\}}{\varepsilon^2}$	Zurek and Chen (2024)	No
$\sqrt{n}$-Horizon DMDP Reduction	$SA\frac{\|h^\star\|_{\text{span}}^2+1}{\varepsilon^2}$	Zurek and Chen (2024)	No
DMDP Reduction + Horizon Calibration	$SA\frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2}$	Our Theorems 1 and 2	No
Span Penalization	$SA\ \inf_{\pi:\ \rho^\pi \text{ constant}}\left\{\frac{\|h^\pi\|_{\text{span}}}{(\rho^\pi-\rho^\star+\varepsilon)^2}\right\}$	Our Theorem 4	No

Table 1: **Algorithms and sample complexity bounds for average reward MDPs** for finding an ε-optimal policy under a generative model (up to log factors). See Appendix A.2 for the definitions of the complexity parameters $D, \tau_{\text{unif}}, \|h^{\widehat{\pi}}\|_{\text{span}}, \|\widehat{\underline{h}}^\star\|_{\text{span}}, \|\underline{h}^{\widehat{\pi}}\|_{\text{span}}$ used in prior work. Note that the diameter D and uniform mixing time $3\tau_{\text{unif}}$ are both upper bounds of $\|h^\star\|_{\text{span}}$ and can be arbitrarily larger than $\|h^\star\|_{\text{span}}$. The parameters $\|h^{\widehat{\pi}}\|_{\text{span}}, \|\widehat{\underline{h}}^\star\|_{\text{span}}, \|\underline{h}^{\widehat{\pi}}\|_{\text{span}}$ are not generally controlled by $\|h^\star\|_{\text{span}}$. The guarantee for our Span Penalization algorithm involves the infimum over all policies π with constant (state-independent) gain ρ^π; see Theorem 4 for an equivalent guarantee in terms of the dataset size.

$\mathcal{S}_1 \cup \mathcal{S}_2$ such that all states in $\mathcal{S}_1$ are transient under all stationary policies and $\mathcal{S}_2$ is communicating. In weakly communicating MDPs, $\rho^\star$ is a constant vector (all entries are equal). All results in this paper assume that P is weakly communicating. While not used in our results, the definitions of the MDP diameter D and uniform mixing time τ_{unif} are given in Appendix A.2 for completeness. We use the standard $\widetilde{O}(\cdot)$ notation to hide logarithmic factors in $S, A, \|h^\star\|_{\text{span}}, 1/\varepsilon$, and $1/\delta$.

3. Main Results

In this section, we present our algorithms and main results. Our algorithms involve the function $\alpha(\delta, n) = 96\sqrt{\log(24SAn^5/\delta)}\log_2(\log_2(n+4))$, which is $\widetilde{O}(1)$; see Remark 1 for its origin.

3.1. Fixed-n Setting

First we consider the setting where the number of samples per state-action pair, n, is fixed. Our objective is to learn a policy with the best possible rate of suboptimality. We refer to this as the *fixed-n setting*. Our Algorithm 1 is based on using the dataset to form an empirical transition kernel $\widehat{P}$ and then computing a near-optimal policy in the DMDP $(\widehat{P}, r)$ for some discount factor γ. The key technique is a method for automatically calibrating γ (equivalently, the effective horizon $\frac{1}{1-\gamma}$): we try multiple values of γ, and for each we compute a near-optimal policy $\widetilde{\pi}_\gamma$ for the DMDP $(\widehat{P}, r, \gamma)$ and a quantity $\widehat{L}(\gamma)$ that lower bounds its gain. We then use the discount factor $\widehat{\gamma}$ that optimizes this lower bound. For computational efficiency, we only need to try $O(\log n)$ values of γ.

Algorithm 1 Lower Bound Maximization via Horizon Calibration

input: Sample size per state-action pair n

1: **for** each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
2: Collect n samples $S^1_{s,a}, \ldots, S^n_{s,a}$ from $P(\cdot \mid s,a)$
3: Form the empirical transition kernel $\widehat{P}(s' \mid s,a) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{S^i_{s,a} = s'\}$, for all $s' \in \mathcal{S}$
4: **end for**
5: Form geometric discount factor range $\mathcal{H} := \{\gamma : \text{there exists an integer } k \text{ such that } \sqrt{n} \le \frac{1}{1-\gamma} = 2^k \le n\}$
6: **for** each discount factor $\gamma \in \mathcal{H}$ **do**
7: Obtain policy $\widetilde{\pi}_\gamma$ and value function $\widetilde{V}_\gamma$ from SOLVEDMDP$(\widehat{P}, r, \gamma, \frac{1}{n})$
8: Compute objective value $\widehat{L}(\gamma) := (1-\gamma)\min_s \widetilde{V}_\gamma(s) - 2\frac{1-\gamma}{n} - \alpha(\delta,n)\sqrt{\frac{\|\widetilde{V}_\gamma\|_{\text{span}} + \frac{3}{n} + 1}{n}}$
9: **end for**
10: Find $\widehat{\gamma} = \text{argmax}_{\gamma \in \mathcal{H}} \widehat{L}(\gamma)$
11: **return** policy $\widehat{\pi} := \widetilde{\pi}_{\widehat{\gamma}}$, gain lower bound $\widehat{\rho} := \max\{\widehat{L}(\widehat{\gamma}), 0\}\mathbf{1}$

Theorem 1 *Suppose P is weakly communicating. For some constant C_3, with probability at least $1-\delta$, the policy $\widehat{\pi}$ and gain lower bound $\widehat{\rho}$ output by Algorithm 1 satisfy (elementwise)*

$$\rho^{\widehat{\pi}} \ge \widehat{\rho} \ge \rho^\star - C_3\alpha(\delta,n)^2\sqrt{\frac{\|h^\star\|_{\text{span}} + 1}{n}}\mathbf{1}.$$

Theorem 1 shows that Algorithm 1 returns a policy $\widehat{\pi}$ with the minimax optimal rate of suboptimality without using any prior knowledge. The algorithm also returns a performance *certificate* $\widehat{\rho}$, which lower-bounds $\rho^{\widehat{\pi}}$ (and $\rho^\star$), and this bound is tight up to an error of $\widetilde{O}(\sqrt{(\|h^\star\|_{\text{span}} + 1)/n})$.

We allow SOLVEDMDP, used in line 7 of Algorithm 1, to be any subroutine for approximately solving the empirical DMDP $(\widehat{P}, r, \gamma)$, and simply require that it returns a deterministic policy $\widetilde{\pi}_\gamma$ and an approximate value function $\widetilde{V}_\gamma$ such that $\widehat{V}_\gamma^{\widetilde{\pi}_\gamma} \ge \widehat{V}_\gamma^\star - \frac{1}{n}\mathbf{1}$ and $\|\widetilde{V}_\gamma - \widehat{V}_\gamma^\star\|_\infty \le \frac{1}{n}$. This can be done for instance using $O\big(\frac{\log(n/(1-\gamma))}{1-\gamma}\big)$ iterations of value iteration.

Algorithm 1 is reminiscent of sample variance penalization (SVP), a statistical learning algorithm which outputs a hypothesis minimizing the empirical risk plus an estimated variance term (Maurer and Pontil, 2009; Duchi and Namkoong, 2019). Here we clarify the connections and differences, which also provides intuitions for our algorithms. First, Algorithm 1 can be understood as controlling not only certain empirical variance (represented by the last term in the definition of $\widehat{L}(\gamma)$ in line 8) but also a certain bias/approximation error due to discounted reduction. This is a significant difference since prior to our work it was not clear that such approximation error could be estimated/controlled without knowing $\|h^\star\|_{\text{span}}$. See our proof sketch in Section 4 for more on this issue. Second, supposing that a lower bound like $\rho^\pi \geq (1-\gamma)\min_s \widehat{V}_\gamma^\pi(s)\mathbf{1} - \sqrt{\|\widehat{V}_\gamma^\pi\|_{\text{span}} + 1/n}\mathbf{1}$ holds for all policies π and all γ with high probability,[1] then an analogue of SVP would be (1) below, while Algorithm 1 can instead be seen as (approximately) solving (2).

$$\max_{\pi,\gamma} (1-\gamma)\min_s \widehat{V}_\gamma^\pi(s) - \sqrt{\frac{\|\widehat{V}_\gamma^\pi\|_{\text{span}} + 1}{n}} \quad (1) \qquad \max_{\gamma} (1-\gamma)\min_s \widehat{V}_\gamma^\star(s) - \sqrt{\frac{\|\widehat{V}_\gamma^\star\|_{\text{span}} + 1}{n}} \quad (2)$$

Since $\max_\pi \widehat{V}_\gamma^\pi = \widehat{V}_\gamma^\star$, solving (2) can be understood as only choosing π to optimize $\widehat{V}_\gamma^\pi$ and then controlling the objective via γ tuning, whereas (1) optimizes all objective terms jointly. While (1) may appear more principled, it is not immediately clear how to solve such a problem, whereas (2) can simply utilize any DMDP solver for optimizing $\widehat{V}_\gamma^\pi$ for fixed γ, then tune γ afterwards. However, this is not the final word on (1), as in Subsection 3.3 we develop our Algorithm 4 based on solving (1).

3.2. Fixed-ε Setting

We next present our Algorithm 2 for the setting where one is given a target suboptimality ε, and the goal is to return a policy with suboptimality bounded by ε using as few samples as possible. We refer to this setting as the *fixed-ε setting*. At a high level, we run our algorithm for the fixed-n setting for a geometrically increasing sequence of dataset sizes $\{n_i\}_i$. However, the fixed-ε setting is more challenging, because beyond lower-bounding the gain of some known policies, to obtain a termination condition we additionally need an (observable) upper bound on the optimal gain $\rho^\star$, meaning we need to bound the gains of all policies. On iteration i with a dataset of size n_i, we compute both lower and upper bounds $\widehat{L}_i(\gamma)$ and $\widehat{U}_i(\gamma)$ for a range of γ, yielding different confidence intervals. The algorithm terminates once one such interval is sufficiently small and certifies the desired suboptimality level ε. We provide more details on how to compute such an upper bound $\widehat{U}_i(\gamma)$ without knowing $\|h^\star\|_{\text{span}}$ in the proof sketches in Section 4.

Theorem 2 *Suppose P is weakly communicating. There exist constants C_1, C_2 such that for any $\varepsilon > 0$, with probability at least $1-\delta$, Algorithm 2 uses at most*

$$N := 4C_1 \frac{\|h^\star\|_{\text{span}} + 1}{\varepsilon^2} \log^3\left(\frac{C_2 SA(\|h^\star\|_{\text{span}} + 1)}{\delta\varepsilon}\right)$$

1. For a single policy π a similar bound follows from our techniques. Uniformity over all π's would incur an additional $\sqrt{S}$.

Algorithm 2 Confidence Interval Minimization via Horizon Calibration

input: Target suboptimality $\varepsilon > 0$

Set iteration number $i = 0$

repeat

  $i \leftarrow i + 1$; set sample size per state-action pair $n_i = 2^i$

  for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**

    Collect n_i samples $S^1_{s,a}, \ldots, S^{n_i}_{s,a}$ from $P(\cdot \mid s, a)$

    Form the ith empirical transition kernel $\widehat{P}^{(i)}(s' \mid s, a) = \frac{1}{n} \sum_{j=1}^{n_i} \mathbb{I}\{S^j_{s,a} = s'\}, \forall s' \in \mathcal{S}$

  end for

  Form geometric discount factor range $\mathcal{H}_i := \{\gamma : \text{there exists an integer } k \text{ such that } \sqrt{n_i} \leq \frac{1}{1-\gamma} = 2^k \leq n_i\}$

  for each discount factor $\gamma \in \mathcal{H}_i$ **do**

    Obtain policy $\widetilde{\pi}_{\gamma,i}$ and value function $\widetilde{V}_{\gamma,i}$ from SOLVEDMDP$(\widehat{P}^{(i)}, r, \gamma, \frac{1}{n_i})$

    Compute upper bound $\widehat{U}_i(\gamma) := (1 - \gamma) \max_s \widetilde{V}_{\gamma,i}(s) + 5\frac{1-\gamma}{n_i} + \frac{2\alpha(\delta,n_i)^2}{(1-\gamma)n_i} + 4\alpha(\delta, n_i)\sqrt{\frac{\|\widetilde{V}_{\gamma,i}\|_{\text{span}} + 1 + \frac{3}{n_i}}{n_i}}$

    Compute lower bound $\widehat{L}_i(\gamma) := (1-\gamma) \min_s \widetilde{V}_{\gamma,i}(s) - 2\frac{1-\gamma}{n_i} - \alpha(\delta, n_i)\sqrt{\frac{\|\widetilde{V}_{\gamma,i}\|_{\text{span}} + \frac{3}{n_i} + 1}{n_i}}$

  end for

  Find discount factor with the smallest interval $\widehat{\gamma}_i := \text{argmin}_{\gamma \in \mathcal{H}_i} \widehat{U}_i(\gamma) - \widehat{L}_i(\gamma)$

until $\widehat{U}_i(\widehat{\gamma}_i) - \widehat{L}_i(\widehat{\gamma}_i) \leq \varepsilon$

return policy $\widehat{\pi} := \widetilde{\pi}_{\widehat{\gamma}_i,i}$, optimal gain upper bound $\widehat{U} := \widehat{U}_i(\widehat{\gamma}_i)$ and lower bound $\widehat{L} := \widehat{L}_i(\widehat{\gamma}_i)$

samples per state-action pair and terminates after at most $\log_2(N)$ *outer iterations. Upon termination Algorithm 2 returns a policy* $\widehat{\pi}$ *and estimates* $\widehat{U}, \widehat{L}$ *such that (elementwise)*

$$\widehat{L}\mathbf{1} \leq \rho^{\widehat{\pi}} \leq \rho^\star \leq \widehat{U}\mathbf{1} \qquad \textit{and} \qquad \widehat{U} - \widehat{L} \leq \varepsilon.$$

In particular, we have

$$\rho^{\widehat{\pi}} \geq \rho^\star - \varepsilon\mathbf{1}.$$

In fact, as shown in Lemma 9, the quantities $\widehat{U}_i(\gamma)$ are valid upper bounds of $\rho^\star$ for all i and $\gamma \in \mathcal{H}_i$, so the algorithm would still be correct if we instead used $\min_{i,\gamma \in \mathcal{H}_i} \widehat{U}_i(\gamma) - \max_{i',\gamma' \in \mathcal{H}_i} \widehat{L}_{i'}(\gamma') \leq \varepsilon$ as a termination condition; that is, we could use a different γ, i to compute the upper bound than the lower bound. The output policy should correspond to the best lower bound, that is, the $\widetilde{\pi}_{\gamma_i,i}$ such that $i, \gamma \in \text{argmax}_{i',\gamma' \in \mathcal{H}_i} \widehat{L}_{i'}(\gamma')$. Also, within each outer-level iteration i of Algorithm 2 we consider $O(\log n_i)$ values of γ, so by Theorem 2 the algorithm terminates after $\widetilde{O}(1)$ total iterations.

Compared to the algorithm of Zurek and Chen (2025), which also takes a discounted reduction approach but uses prior knowledge of $\|h^\star\|_{\text{span}}$ to set the discount factor, the $\widehat{\gamma}_i$ chosen by our algorithm should not be seen as implicitly estimating $\|h^\star\|_{\text{span}}$ and in fact cannot be used to do so (consistent with known hardness results on span estimation (Tuynman et al., 2024; Zurek and Chen, 2025)). Rather, $\widehat{\gamma}_i$ can be understood as balancing bounds on certain approximation and estimation error terms, potentially in a superior way than the $\|h^\star\|_{\text{span}}$-knowledge-based choice

for non-worst-case instances. See Appendix B for further discussion of the relationship between $\|h^\star\|_{\text{span}}$ and the $\widehat{\gamma}_i$ computed in each iteration of our algorithm.

3.3. Span Penalization and Oracle Inequalities

Finally, we return to the fixed-n setting and the goal of implementing the formulation (1) that resembles sample variance penalization. As we see momentarily, doing so allows us to obtain a stronger "oracle inequality" that optimally trades off suboptimality and complexity.

By superfluously introducing a span constraint of the form $\|\widehat{V}_\gamma^\pi\|_{\text{span}} \le M$ to (1) and optimizing over M, we obtain the equivalent optimization problem

$$(1) \;\equiv\; \max_{\gamma, M} \; \max_{\pi : \|\widehat{V}_\gamma^\pi\|_{\text{span}} \le M} (1-\gamma) \min_s \widehat{V}_\gamma^\pi(s) - \sqrt{\frac{M+1}{n}}.$$

These manipulations are useful if, for fixed (γ, M), we are able to solve the span-constrained optimization problem $\max_{\pi : \|\widehat{V}_\gamma^\pi\|_{\text{span}} \le M} \widehat{V}_\gamma^\pi$. A natural approach is to attempt to apply value iteration with a span truncation step. This is inspired by Fruit et al. (2018), who first introduced this truncation operator and combined it with average-reward (undiscounted) value iteration to solve a certain bias-constrained gain optimization problem. Thus we define the span truncation operator $\text{Clip}_M : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ where

$$\text{Clip}_M(V)(s) := \begin{cases} V(s) & \text{if } V(s) \le M + \min_{s'} V(s') \\ M + \min_{s'} V(s') & \text{otherwise,} \end{cases} \tag{3}$$

or equivalently $\text{Clip}_M(V) = \min\{V, (M + \min_{s'} V(s'))\mathbf{1}\}$, where the outer $\min$ is elementwise. By combining Clip_M, which is $\|\cdot\|_\infty$-nonexpansive, with the discounted Bellman operator $\mathcal{T}_\gamma(V) := M(r + \gamma PV)$, we can now define our *Span-Constrained Planning* subroutine, given as Algorithm 3.

Algorithm 3 Span-Constrained Planning

input: Discounted MDP (P, r, γ), span constraint bound $M > 0$, target error $\varepsilon > 0$

1: Form clipped discounted Bellman operator $\mathcal{L} := \text{Clip}_M \circ \mathcal{T}_\gamma$
2: Set initial point $V^0 = \mathbf{0} \in \mathbb{R}^{\mathcal{S}}$, total iteration count $T = \left\lceil \frac{\log(\frac{3}{(1-\gamma)^2 \varepsilon})}{1-\gamma} \right\rceil$
3: **for** $t = 0, \dots, T-1$ **do**
4: $\quad V^{t+1} = \mathcal{L}(V^t)$
5: **end for**
6: Set $\widehat{\pi}(s) \in \text{argmax}_{a \in \mathcal{A}} r(s,a) + \gamma P_{sa} V^T$ for all $s \in \mathcal{S}$
7: Compute minimum state value $m = \min_{s \in \mathcal{S}} V^T(s)$
8: Set truncated reward $\tilde{r}(s,a) = \min\{m + M - \gamma P_{sa} V^T, r(s,a)\}$ for all $s \in \mathcal{S}, a \in \mathcal{A}$
9: **return** policy $\widehat{\pi}$, approximate value function V^T, truncated reward $\tilde{r}$

Now we discuss why this subroutine returns a truncated reward $\tilde{r}$ and whether it solves the aforementioned span-constrained planning problem $\max_{\pi : \|V_\gamma^\pi\|_{\text{span}} \le M} V_\gamma^\pi$. (Here we discuss value functions based on a generic P rather than $\widehat{P}$ for clarity.) While the clipped Bellman operator $\mathcal{L} = \text{Clip}_M \circ \mathcal{T}_\gamma$ has a unique fixed point $V_{\gamma,M}^\star$, this fixed point generally may not satisfy any Bellman equation $V_{\gamma,M}^\star = M^\pi(r + \gamma P V_{\gamma,M}^\star)$ for any policy π. This is to be expected, because such a

Bellman equation would imply that policy π has value function $V_\gamma^\pi = V_{\gamma,M}^\star$ and thus $\|V_\gamma^\pi\|_{\text{span}} \le M$, but it is possible that no policies have value functions with span $\le M$. (For example consider an MDP with $A = 1$ where the only policy has a value function with large span.) The truncated reward $\tilde{r}$ remedies both of these closely related issues: it is defined to (approximately) satisfy a Bellman equation $V_{\gamma,M}^\star = M^{\widehat{\pi}}(\tilde{r} + \gamma P V_{\gamma,M}^\star)$ for some $\widehat{\pi}$, which would then imply that in the DMDP $(P, \tilde{r}, \gamma)$ with the truncated reward, the value function $V_{\gamma,\tilde{r}}^{\widehat{\pi}}$ of policy $\widehat{\pi}$ does indeed have $\|V_{\gamma,\tilde{r}}^{\widehat{\pi}}\|_{\text{span}} \le M$. Thus, we do not actually solve $\max_{\pi:\|V_\gamma^\pi\|_{\text{span}} \le M} V_\gamma^\pi$, but instead the relaxed problem $\max_{\pi,\tilde{r}:\|V_{\gamma,\tilde{r}}^\pi\|_{\text{span}} \le M, \tilde{r} \le r} V_{\gamma,\tilde{r}}^\pi$, which leads to a value function $V_{\gamma,\tilde{r}}^{\widehat{\pi}}$ that is at least as large as $\max_{\pi:\|V_\gamma^\pi\|_{\text{span}} \le M} V_\gamma^\pi$ yet still has span bounded by M. And, since elementwise $\tilde{r} \le r$, the actual value function of $\widehat{\pi}$, $V_\gamma^{\widehat{\pi}}$, is lower-bounded by $V_{\gamma,\tilde{r}}^{\widehat{\pi}}$, which is still compatible with the lower-bound-based approach taken in Algorithm 4, which we are now prepared to develop.

The key properties of the Span-Constrained Planning Algorithm 3 are summarized in the following lemma.

Lemma 3 *The operator $\mathcal{L}$ defined in Algorithm 3 is γ-contractive and has unique fixed point $V_{\gamma,M}^\star$. Moreover, the $\widehat{\pi}, V^T$, and $\tilde{r}$ returned by Algorithm 3 satisfy*

1. *(proximity to exact fixed point) $\|V^T - V_{\gamma,M}^\star\|_\infty \le \varepsilon$;*
2. *(near-feasibility of $\tilde{r}$ and $\widehat{\pi}$) $\tilde{r} \le r$, $\|V_{\gamma,\tilde{r}}^{\widehat{\pi}} - V_{\gamma,M}^\star\|_\infty \le \varepsilon$, and $\|V_{\gamma,\tilde{r}}^{\widehat{\pi}}\|_{\text{span}} \le M + 2\varepsilon$;*
3. *(near-optimality of $\widehat{\pi}$) for any policy π' and reward function $r' \le r$ such that $\|V_{\gamma,r'}^{\pi'}\|_{\text{span}} \le M$, we have $V_{\gamma,M}^\star \ge V_{\gamma,r'}^{\pi'}$, and $V_\gamma^{\widehat{\pi}} \ge V_{\gamma,\tilde{r}}^{\widehat{\pi}} \ge V_{\gamma,M}^\star - \varepsilon\mathbf{1} \ge V_{\gamma,r'}^{\pi'} - \varepsilon\mathbf{1}$.*

We can now present our final Algorithm 4, titled *Empirical Span Penalization*. In summary, it can be understood as solving the problem

$$\max_{\gamma,M} \max_{\pi,\tilde{r}:\|\widehat{V}_{\gamma,\tilde{r}}^\pi\|_{\text{span}} \le M, \tilde{r} \le r} \min_s \ \widehat{V}_{\gamma,\tilde{r}}^\pi(s) - \sqrt{\frac{M+1}{n}}.$$

For reasons apparent in the proof sketches in Section 4, it is unclear whether there is any provable benefit of optimizing over all γ as opposed to choosing γ such that $\frac{1}{1-\gamma} \approx \sqrt{nM}$, so we opt to only optimize M and simply set γ in this way. Like for our previous algorithms, we only need to try $O(\log n)$ values of M to approximately solve the problem.

Theorem 4 *Suppose P is weakly communicating. For some constant C_4, with probability at least $1 - \delta$, the policy $\widehat{\pi}$ and gain lower bound $\widehat{\rho}$ output by Algorithm 4 satisfy (elementwise)*

$$\rho^{\widehat{\pi}} \ge \widehat{\rho} \ge \sup_{\pi:\rho^\pi \text{ constant}} \left\{ \rho^\pi - C_4 \alpha(\delta, n) \sqrt{\frac{\|h^\pi\|_{\text{span}} + 1}{n}} \mathbf{1} \right\}.$$

Theorem 4 shows that the output policy $\widehat{\pi}$ satisfies an "oracle inequality": $\widehat{\pi}$ competes with the constant-gain policy π that has the best tradeoff between suboptimality and complexity, as measured by $\rho^\star - \rho^{\widehat{\pi}}$ and $\|h^\pi\|_{\text{span}}$, respectively. This bound, when written in the equivalent form

$$\rho^\star - \rho^{\widehat{\pi}} \le \inf_{\pi:\rho^\pi \text{ constant}} \left\{ (\rho^\star - \rho^\pi) + C_4 \alpha(\delta, n) \sqrt{\frac{\|h^\pi\|_{\text{span}} + 1}{n}} \mathbf{1} \right\},$$

Algorithm 4 Empirical Span Penalization

input: Sample size per state-action pair n

for each state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
  Collect n samples $S^1_{s,a}, \dots, S^n_{s,a}$ from $P(\cdot \mid s,a)$
  Form the empirical transition kernel $\widehat{P}(s' \mid s,a) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{S^i_{s,a} = s'\}$, for all $s' \in \mathcal{S}$
end for
for $i = 2, \dots, \lceil \log_2 n \rceil$ **do**
  Set span constraint bound $M_i = 2^i$
  Set discount factor γ_i such that $\frac{1}{1-\gamma_i} = \max\left\{\frac{\sqrt{nM_i}}{\alpha(\delta,n)2\sqrt{2}}, 1\right\}$
  Obtain policy $\widetilde{\pi}_i$, value function $\widetilde{V}_i$, truncated reward $\tilde{r}_i$ from the Span-Constrained Planning Algorithm 3 with input DMDP $(\widehat{P}, r, \gamma_i)$, span constraint bound M_i, and target error $\frac{1}{n}$
  Compute objective value $\widehat{L}(i) := (1-\gamma_i)\min_s \widetilde{V}_i(s) - \alpha(\delta,n)\sqrt{\frac{M_i+1}{n}}$
end for
Find $\widehat{i} = \operatorname{argmax}_i \widehat{L}(i)$
return policy $\widehat{\pi} := \widetilde{\pi}_{\widehat{i}}$, gain lower bound $\widehat{\rho} := \max\{\widehat{L}(\widehat{i}), 0\}\mathbf{1}$

resembles the oracle inequalities (Deheuvels et al., 2007; Koltchinskii, 2011) from the statistics literature that feature a similar tradeoff.[2] It is immediate that Theorem 4 is at least as strong as Theorem 1 by setting π to be the Blackwell-optimal policy $\pi^\star$, since $\rho^{\pi^\star} = \rho^\star$ is constant and $\|h^{\pi^\star}\|_{\text{span}} = \|h^\star\|_{\text{span}}$ by definition.

In fact, since there may generally be many gain-optimal π's with $h^\pi \neq h^\star$, Theorem 4 implies

$$\rho^{\widehat{\pi}} \geq \rho^\star - C_4\alpha(\delta,n)\sqrt{\frac{\inf_{\pi:\rho^\pi=\rho^\star}\|h^\pi\|_{\text{span}} + 1}{n}}\mathbf{1},$$

thus replacing $\|h^\star\|_{\text{span}}$ with the smallest bias of any gain-optimal policy. The above bound does not contradict the minimax lower bound rate of $\sqrt{\|h^\star\|_{\text{span}}/n}$ (Wang et al., 2022), which is based on worst-case instances, but we present an example in Appendix D where $\inf_{\pi:\rho^\pi=\rho^\star}\|h^\pi\|_{\text{span}}$ can be arbitrarily smaller than $\|h^\star\|_{\text{span}}$. To the best of our knowledge, no other algorithm can achieve such a bound. Another situation where Theorem 4 outperforms the minimax rate is when all gain-optimal policies have large span relative to n but there is some near-optimal policy with much smaller span, such that learning a near-optimal policy is still possible. We provide such an example in Appendix D. In practice, RL is commonly applied to problems for which exact optimal policies are extremely complicated, and yet such problems may be solved to a reasonable degree of optimality by following relatively simple heuristics. Our result shows that it is possible to be automatically adaptive to the policy with the best tradeoff of complexity and suboptimality.

4. Proof Sketches

In this section, we outline the key ideas in the proofs of our main theorems.

2. An oracle inequality controls the error of a statistical estimator in terms of that of an oracle that selects the optimal model by trading off approximation error and estimation error.

4.1. Proof sketch for Theorem 1

We start by outlining the ideas behind Algorithm 1 and Theorem 1 for the fixed-n setting. First we review the approach of Zurek and Chen (2025) and simplifications due to Zurek and Chen (2024), which achieve the same rate as Theorem 1 but require knowing $\|h^\star\|_{\text{span}}$. The algorithm of Zurek and Chen (2025) chooses a discount factor $\gamma^\star$ in a way that depends crucially on $\|h^\star\|_{\text{span}}$, and then solves the $\gamma^\star$-discounted empirical MDP $(\widehat{P}, r, \gamma^\star)$, where $\widehat{P}$ is constructed as in Algorithm 1. The resulting policy $\widehat{\pi}_{\gamma^\star}$ is shown to be near-optimal for the AMDP (P, r) using the reduction from Wang et al. (2022). We now illustrate how this optimally-chosen $\gamma^\star$ trades off certain approximation and learning error terms. Letting $\widehat{\pi}_\gamma$ be the policy output by the above-described procedure but with a general discount factor γ, it is shown that

$$\begin{aligned}\left\|\rho^{\widehat{\pi}_\gamma} - \rho^\star\right\|_\infty &\overset{(i)}{\lesssim} (1-\gamma)\left(\|h^\star\|_{\text{span}} + \|V_\gamma^{\widehat{\pi}_\gamma} - V_\gamma^\star\|_\infty\right) \\ &\overset{(ii)}{\lesssim} (1-\gamma)\left(\|h^\star\|_{\text{span}} + \frac{1}{1-\gamma}\sqrt{\frac{\|V_\gamma^\star\|_{\text{span}} + \|\widehat{V}_\gamma^\star\|_{\text{span}} + 1}{n}}\right) \\ &\overset{(iii)}{\lesssim} (1-\gamma)\|h^\star\|_{\text{span}} + \sqrt{\frac{\|h^\star\|_{\text{span}} + \|\widehat{V}_\gamma^\star\|_{\text{span}} + 1}{n}}, \end{aligned} \tag{4}$$

where the notation $\lesssim$ ignores constant/log factors, (i) is the AMDP-to-DMDP reduction of Wang et al. (2022), (ii) uses Zurek and Chen (2024, Theorem 9) to bound $\|V_\gamma^{\widehat{\pi}_\gamma} - V_\gamma^\star\|_\infty$ in terms of the variance parameters $\|V_\gamma^\star\|_{\text{span}}, \|\widehat{V}_\gamma^\star\|_{\text{span}}$, and (iii) uses $\|V_\gamma^\star\|_{\text{span}} \lesssim \|h^\star\|_{\text{span}}$ due to Wei et al. (2020, Lemma 2). To proceed, one can invoke Zurek and Chen (2024, Lemma 12) (itself summarizing a key step due to Zurek and Chen 2025 similar to our Lemma 7), which controls the variance parameter $\|\widehat{V}_\gamma^\star\|_{\text{span}}$ associated with the empirically optimal policy $\widehat{\pi}$ like

$$\left\|\widehat{V}_\gamma^\star\right\|_{\text{span}} \lesssim \left\|V_\gamma^\star\right\|_{\text{span}} + \left\|\widehat{V}_\gamma^\star - V_\gamma^\star\right\|_\infty \lesssim \|h^\star\|_{\text{span}} + \frac{1}{1-\gamma}\sqrt{\frac{\|h^\star\|_{\text{span}} + \|\widehat{V}_\gamma^\star\|_{\text{span}} + 1}{n}}. \tag{5}$$

Solving this recursion yields $\|\widehat{V}_\gamma^\star\|_{\text{span}} \lesssim \|h^\star\|_{\text{span}} + \frac{1}{(1-\gamma)^2 n}$. Plugging back into (4) gives the following bound and in particular the term T_3:

$$\left\|\rho^{\widehat{\pi}_\gamma} - \rho^\star\right\|_\infty \lesssim \underbrace{(1-\gamma)\|h^\star\|_{\text{span}}}_{T_1} + \underbrace{\sqrt{\frac{\|h^\star\|_{\text{span}} + 1}{n}}}_{T_2} + \underbrace{\frac{1}{(1-\gamma)n}}_{T_3} \tag{6}$$

Intuitively, this argument is reminiscent of localization techniques in statistical learning (e.g., Bartlett et al. 2005) since it controls the variance parameters associated with near-optimal policies in terms of the variance of an optimal policy. For this reason we call T_3 the *localization error*. Term T_1 is the approximation error due to the DMDP reduction, and term T_2 is independent of γ and is exactly the desired minimax rate. From (6) it is immediate that choosing $\gamma^\star$ so that $\frac{1}{1-\gamma^\star} \approx \sqrt{n(\|h^\star\|_{\text{span}} + 1)}$ will balance all three terms and achieve the minimax rate.

Now we derive an algorithm which does not require prior knowledge of $\|h^\star\|_{\text{span}}$. The simplest idea is to attempt to estimate $\|h^\star\|_{\text{span}}$ and plug the estimate into the formula for $\gamma^\star$. However,

estimating $\|h^\star\|_{\text{span}}$ up to a constant factor is shown to be impossible in the worst case (Zurek and Chen, 2025; Tuynman et al., 2024). Moreover, Zurek and Chen (2024, Theorem 14) provides an example where choosing γ too large provably fails to obtain the minimax-optimal rate, suggesting that the localization error term T_3 could not be removed with a different analysis. A partially successful approach is to attempt to replace the terms in (6) with *observable/estimable* upper bounds that depend on γ, because observability would enable us to minimize these upper bounds over γ. Doing a similar "localization" approach as in (5), but instead upper-bounding $\|V_\gamma^\star\|_{\text{span}}$ in terms of $\|\widehat{V}_\gamma^\star\|_{\text{span}}$, one obtains $\|V_\gamma^\star\|_{\text{span}} \lesssim \|\widehat{V}_\gamma^\star\|_{\text{span}} + \frac{1}{(1-\gamma)^2 n}$. Using this bound but otherwise following the derivation of (6), we can get

$$\left\|\rho^{\widehat{\pi}_\gamma} - \rho^\star\right\|_\infty \lesssim \underbrace{(1-\gamma)\|h^\star\|_{\text{span}}}_{T_1} + \underbrace{\sqrt{\frac{\|\widehat{V}_\gamma^\star\|_{\text{span}} + 1}{n}}}_{T_2'} + \underbrace{\frac{1}{(1-\gamma)n}}_{T_3}, \tag{7}$$

where now both T_3 and T_2' are observable, and based on (5), there are some choices of γ (including $\gamma^\star$) for which T_2' will not be too much larger than T_2. However, T_1 still depends on the unknown $\|h^\star\|_{\text{span}}$. This approach can actually be salvaged with one key modification and forms the basis of our result (Theorem 2), but here we take a slightly different approach to address term T_1. Since T_1 is decreasing monotonically in γ, we believe *Lepski's trick* (Lepski and Spokoiny, 1997) could actually be applied to estimate $\rho^\star$ with the minimax rate, but obtaining a policy with a matching order of suboptimality appears to require more work.

In particular, instead of optimizing a bound on the suboptimality $\|\rho^{\widehat{\pi}_\gamma} - \rho^\star\|_\infty$ of the policy $\widehat{\pi}_\gamma$ that is optimal for the empirical DMDP $(\widehat{P}, r, \gamma)$, we optimize a lower bound on the performance of $\widehat{\pi}_\gamma$ since we are considering the fixed-n setting. For any π and any γ, it is known that $\rho^\pi \geq (1-\gamma)V_\gamma^\pi - (1-\gamma)\|V_\gamma^\pi\|_{\text{span}}\mathbf{1}$.[3] The significance of this statement applied to our problem is that if we only desire a lower bound for $\rho^{\widehat{\pi}_\gamma}$, then the approximation error term T_1 can be replaced with $(1-\gamma)\|V_\gamma^{\widehat{\pi}_\gamma}\|_{\text{span}}$, which can be estimated. Specifically, using the bound $\|V_\gamma^{\widehat{\pi}_\gamma} - \widehat{V}_\gamma^\star\|_\infty \lesssim \frac{1}{1-\gamma}\sqrt{\frac{\|\widehat{V}_\gamma^\star\|_{\text{span}}+1}{n}}$, we have

$$\begin{aligned}\rho^{\widehat{\pi}_\gamma} &\geq (1-\gamma)V_\gamma^{\widehat{\pi}_\gamma} - (1-\gamma)\|V_\gamma^{\widehat{\pi}_\gamma}\|_{\text{span}}\mathbf{1} \geq (1-\gamma)\widehat{V}_\gamma^\star - (1-\gamma)\left(\|\widehat{V}_\gamma^\star\|_{\text{span}} - C\|V_\gamma^{\widehat{\pi}_\gamma} - \widehat{V}_\gamma^\star\|_\infty\right)\mathbf{1} \\ &\geq (1-\gamma)\widehat{V}_\gamma^\star - (1-\gamma)\|\widehat{V}_\gamma^\star\|_{\text{span}}\mathbf{1} - C'\sqrt{\frac{\|\widehat{V}_\gamma^\star\|_{\text{span}} + 1}{n}}\mathbf{1}, \end{aligned} \tag{8}$$

where $C, C' \leq \widetilde{O}(1)$. Let $L(\gamma)$ be the minimum entry of the RHS of (8). Then $L(\gamma)$ is observable and we can show $L(\gamma^\star) \geq \rho^\star - \sqrt{\frac{\|h^\star\|_{\text{span}}+1}{n}}$, so returning the policy $\pi_{\widehat{\gamma}}$ where $\widehat{\gamma} = \text{argmax}_\gamma L(\gamma)$, we obtain the minimax rate. (No explicit localization error term appears in (8), but lower-bounding $L(\gamma)$ requires the "localization bound" $\|\widehat{V}_\gamma^\star\|_{\text{span}} \lesssim \|h^\star\|_{\text{span}} + \frac{1}{(1-\gamma)^2 n}$, so it still appears implicitly in the size of $\|\widehat{V}_\gamma^\star\|_{\text{span}}$.) This is essentially our Theorem 1.

4.2. Proof sketch for Theorem 2

Next we describe the ideas behind Algorithm 2 and Theorem 2 for the fixed-ε setting. It suffices to develop a method of bounding the suboptimality $\|\rho^{\widehat{\pi}} - \rho^\star\|_\infty$ within the fixed-n setting, as we

3. We actually refine this slightly to $\rho^\pi \geq (1-\gamma)\min_s V_\gamma^\pi(s)\mathbf{1}$, which is what appears in our algorithms.

can then double the dataset size until the suboptimality bound is $\leq \varepsilon$. We return to (7) and the problematic term T_1, which originates from the AMDP-to-DMDP reduction of Wang et al. (2022). We show that actually $\|h^\star\|_{\text{span}}$ can be replaced with $\|V_\gamma^\star\|_{\text{span}}$, that is,

$$\left\|\rho^{\widehat{\pi}_\gamma} - \rho^\star\right\|_\infty \lesssim (1-\gamma)\left(\|V_\gamma^\star\|_{\text{span}} + \left\|V_\gamma^{\widehat{\pi}_\gamma} - V_\gamma^\star\right\|_\infty\right). \tag{9}$$

See Lemma 5 for the precise statement. We could essentially recover the reduction result of Wang et al. (2022) (step (i) of (6)) by using $\|V_\gamma^\star\|_{\text{span}} \lesssim \|h^\star\|_{\text{span}}$, but the subtle improvement of (9) gives a new approximation error term $(1-\gamma)\|V_\gamma^\star\|_{\text{span}}$ which, with aforementioned arguments, can be bounded by fully observable quantities involving $\widehat{V}_\gamma^\star$. Specifically, (7) can be replaced with

$$\left\|\rho^{\widehat{\pi}_\gamma} - \rho^\star\right\|_\infty \lesssim (1-\gamma)\|\widehat{V}_\gamma^\star\|_{\text{span}} + \sqrt{\frac{\|\widehat{V}_\gamma^\star\|_{\text{span}} + 1}{n_i}} + \frac{1}{(1-\gamma)n_i}, \tag{10}$$

where we now use a variable dataset size n_i, since we intend to double the size until a termination condition is satisfied. Comparing with our Algorithm 2, the RHS of (10) is essentially the difference between the upper and lower bounds defined in the algorithm, that is the quantity $\widehat{U}_i(\gamma) - \widehat{L}_i(\gamma)$. Thus Algorithm 2 is slightly more useful than the procedure sketched here, since it provides valid upper and lower bounds rather than just the difference between such bounds, but this does not require much more work. Similarly to the previous algorithm, when $\gamma = \gamma^\star$, the bound (10) will be $\lesssim \sqrt{(\|h^\star\|_{\text{span}} + 1)/n}$ thanks to the "localization bound" $\|\widehat{V}_\gamma^\star\|_{\text{span}} \lesssim \|h^\star\|_{\text{span}} + \frac{1}{(1-\gamma)^2 n}$ arising from (5). This concludes our motivation and proof sketch of Theorem 2. Also see Appendix B for more discussion on the relationship between $\|h^\star\|_{\text{span}}$ and the $\widehat{\gamma}_i$ that minimizes the right hand side of (10) over γ.

4.3. Proof sketch of Theorem 4

For Theorem 4, our starting point is a lower bound similar to (8) used within Algorithm 1, but with a few key differences. Suppose we apply the Span-Constrained Planning Algorithm 3 to the empirical DMDP $(\widehat{P}, r, \gamma)$ with span constraint M and, for simplicity, an arbitrarily small target error, to get the policy $\widehat{\pi}_M$, truncated reward $\tilde{r}$, and the exact fixed point $\widehat{V}_{\gamma,M}^\star = \widehat{V}_{\gamma,\tilde{r}}^{\widehat{\pi}_M}$. We also fix a comparator policy π such that $\|h^\pi\|_{\text{span}} + 1 \leq \frac{M}{4}$ and ρ^π is constant, which enable us to show that $\|V_\gamma^\pi\|_{\text{span}} \lesssim \|h^\pi\|_{\text{span}}$ for any γ and that $\|\widehat{V}_\gamma^\pi - V_\gamma^\pi\|_\infty \lesssim \sqrt{M/(1-\gamma)^2 n}$ using standard techniques (Zurek and Chen, 2024). By setting γ so that $\frac{1}{1-\gamma} \lesssim \sqrt{nM}$, we can ensure that $\|\widehat{V}_\gamma^\pi\|_{\text{span}} \leq M$ (by bounding it in terms of $\|V_\gamma^\pi\|_{\text{span}} \lesssim \|h^\pi\|_{\text{span}} \lesssim M$ and $\|\widehat{V}_\gamma^\pi - V_\gamma^\pi\|_\infty$), which is essential to ensure that π is "feasible" for the empirical span-constrained problem, guaranteeing that $\widehat{V}_{\gamma,\tilde{r}}^{\widehat{\pi}_M} \geq \widehat{V}_\gamma^\pi$ by Lemma 3. These assumptions and condition on γ also imply that $\widehat{V}_\gamma^\pi \geq \frac{1}{1-\gamma}\rho^\pi - O(M)$. Combining

all these steps we can show that

$$\begin{aligned}\rho^{\widehat{\pi}_M} &\geq (1-\gamma)\min_s V_\gamma^{\widehat{\pi}_M}(s)\mathbf{1} \overset{\text{(I)}}{\geq} (1-\gamma)\min_s V_{\gamma,\tilde{r}}^{\widehat{\pi}_M}(s)\mathbf{1} \\ &\geq (1-\gamma)\min_s \widehat{V}_{\gamma,\tilde{r}}^{\widehat{\pi}_M}(s)\mathbf{1} - (1-\gamma)\left\|\widehat{V}_{\gamma,\tilde{r}}^{\widehat{\pi}_M} - V_{\gamma,\tilde{r}}^{\widehat{\pi}_M}\right\|_\infty \mathbf{1} \\ &\overset{\text{(II)}}{\geq} (1-\gamma)\min_s \widehat{V}_{\gamma,\tilde{r}}^{\widehat{\pi}_M}(s)\mathbf{1} - C\sqrt{\frac{M+1}{n}} \overset{\text{(III)}}{\geq} (1-\gamma)\min_s \widehat{V}_\gamma^{\pi}(s)\mathbf{1} - C\sqrt{\frac{M+1}{n}} \\ &\overset{\text{(IV)}}{\geq} \rho^\pi - (1-\gamma)C'M\mathbf{1} - C\sqrt{\frac{M+1}{n}},\end{aligned}$$

where $C, C' \leq \widetilde{O}(1)$. Here (I) follows from $\tilde{r} \leq r$, and (II) follows from the error bound

$$\left\|\widehat{V}_{\gamma,\tilde{r}}^{\widehat{\pi}_M} - V_{\gamma,\tilde{r}}^{\widehat{\pi}_M}\right\|_\infty \lesssim \frac{1}{1-\gamma}\sqrt{\frac{\|\widehat{V}_{\gamma,\tilde{r}}^{\widehat{\pi}_M}\|_{\text{span}} + 1}{n}} = \frac{1}{1-\gamma}\sqrt{\frac{M+1}{n}}, \tag{11}$$

and (III) and (IV) follow from the aforementioned "feasibility" of π and the lower bound on $\widehat{V}_\gamma^\pi$. The key new challenge is to develop the error bound (11), which is complicated due to the statistical dependence between $\widehat{P}$ and $\widehat{\pi}_M$. Without the span constraint this is addressed by Agarwal et al. (2020); Li et al. (2020) through the use of "absorbing MDP" constructions, which enable concentration inequalities for $|(\widehat{P}_{sa} - P_{sa})\widehat{V}_\gamma^\star|$. Instead, we desire such bounds for $|(\widehat{P}_{sa} - P_{sa})\widehat{V}_{\gamma,M}^\star|$, that is, involving the empirical span-constrained optimal value functions. Based on the contractivity of the span-constrained Bellman operator $\mathcal{L}$, we develop a new absorbing MDP construction for span-constrained value functions, ultimately leading to the desired bound (11).

5. Conclusion

We resolve the problem of achieving the minimax optimal span-based complexity in average-reward RL without prior knowledge of the span or other unknown MDP complexity parameters. Our algorithms apply to both the fixed-sample-size and fixed-suboptimality settings, and moreover achieve optimal tradeoff between complexity and suboptimality, surpassing the minimax lower bound in benign settings. Future directions of interest include generalizing to general/multichain MDPs and extending beyond the tabular simulator setting.

Acknowledgments

Y. Chen and M. Zurek acknowledge support by National Science Foundation grants CCF-2233152 and DMS-2023239.

References

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1707–1714, 2012.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91 (3):325–349, June 2013. ISSN 1573-0565. doi: 10.1007/s10994-013-5368-1. URL https://doi.org/10.1007/s10994-013-5368-1.

Peter Bartlett and Ambuj Tewari. REGAL: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Uncertainty in Artificial Intelligence: Proceedings of the 25th Conference*, pages 35–42. AUAI Press, 2009.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, August 2005. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053605000000282. URL https://projecteuclid.org/journals/annals-of-statistics/volume-33/issue-4/Local-Rademacher-complexities/10.1214/009053605000000282.full. Publisher: Institute of Mathematical Statistics.

Victor Boone and Zihan Zhang. Achieving Tractable Minimax Optimal Regret in Average Reward MDPs, June 2024. URL http://arxiv.org/abs/2406.01234. arXiv:2406.01234 [cs].

Paul Deheuvels, Eustasio Del Barrio, and Sara Van De Geer. *Lectures on Empirical Processes: Theory and Statistical Applications*, volume 6 of *EMS Series of Lectures in Mathematics*. EMS Press, 1 edition, January 2007. ISBN 978-3-03719-027-2 978-3-03719-527-7. doi: 10.4171/027. URL https://ems.press/doi/10.4171/027.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(1):2450–2504, January 2019. ISSN 1532-4435.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1578–1586. PMLR, 2018.

Ying Jin, Ramki Gummadi, Zhengyuan Zhou, and Jose Blanchet. Feasible Q-Learning for Average Reward Reinforcement Learning. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 1630–1638. PMLR, April 2024. URL https://proceedings.mlr.press/v238/jin24b.html. ISSN: 2640-3498.

Yujia Jin and Aaron Sidford. Efficiently solving MDPs with stochastic mirror descent. In *International Conference on Machine Learning*, pages 4890–4900. PMLR, 2020.

Yujia Jin and Aaron Sidford. Towards tight bounds on the sample complexity of average-reward MDPs. In *International Conference on Machine Learning*, pages 5055–5064. PMLR, 2021.

Michael Kearns and Satinder Singh. Finite-Sample Convergence Rates for Q-Learning and Indirect Algorithms. In *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998. URL https://proceedings.neurips.cc/paper/1998/hash/99adff456950dd9629a5260c4de21858-Abstract.html.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-22146-0 978-3-642-22147-7. doi: 10.1007/978-3-642-22147-7. URL https://link.springer.com/10.1007/978-3-642-22147-7.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, Cambridge ; New York, NY, 2020. ISBN 978-1-108-57140-1.

Oleg V. Lepski and Vladimir G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 1997.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model. In *Advances in Neural Information Processing Systems*, volume 33, pages 12861–12872. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/96ea64f3a1aa2fd00c72faacf0cb8ac9-Abstract.html.

Tianjiao Li, Feiyang Wu, and Guanghui Lan. Stochastic first-order methods for average-reward markov decision processes. *Mathematics of Operations Research*, 2024.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample Variance Penalization, July 2009. URL http://arxiv.org/abs/0907.3740. Proc. Computational Learning Theory Conference (COLT 2009). arXiv:0907.3740 [stat] version: 1.

Gergely Neu and Nneka Okolo. Dealing with unbounded gradients in stochastic saddle-point optimization, June 2024. URL http://arxiv.org/abs/2402.13903. Forty-first International Conference on Machine Learning. arXiv:2402.13903 [cs, math, stat] version: 2.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1 edition, April 1994. ISBN 978-0-471-61977-2 978-0-470-31688-7. doi: 10.1002/9780470316887. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316887.

Adrienne Tuynman, Rémy Degenne, and Emilie Kaufmann. Finding good policies in average-reward Markov Decision Processes without prior knowledge, May 2024. URL http://arxiv.org/abs/2405.17108. arXiv:2405.17108 [cs].

Jinghan Wang, Mengdi Wang, and Lin F. Yang. Near Sample-Optimal Reduction-based Policy Learning for Average Reward MDP, December 2022. URL http://arxiv.org/abs/2212.00603. arXiv:2212.00603 [cs].

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal Sample Complexity of Reinforcement Learning for Mixing Discounted Markov Decision Processes, September 2023. URL http://arxiv.org/abs/2302.07477. arXiv:2302.07477.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR, 2020.

Zihan Zhang and Qiaomin Xie. Sharper Model-free Reinforcement Learning for Average-reward Markov Decision Processes, June 2023. URL http://arxiv.org/abs/2306.16394. The Thirty Sixth Annual Conference on Learning Theory. arXiv:2306.16394 [cs].

Matthew Zurek and Yudong Chen. The Plug-in Approach for Average-Reward and Discounted MDPs: Optimal Sample Complexity Analysis, October 2024. URL http://arxiv.org/abs/2410.07616. International Conference on Algorithmic Learning Theory (ALT 2025). arXiv:2410.07616 [cs].

Matthew Zurek and Yudong Chen. Span-Based Optimal Sample Complexity for Weakly Communicating and General Average Reward MDPs. *Advances in Neural Information Processing Systems*, 37:33455–33504, January 2025. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3acbe9dc3a1e8d48a57b16e9aef91879-Abstract-Conference.html.

Appendix A. Additional Notation and Guide to Appendices

In this section, we provide additional notations and definitions, and outline the organization of the remainder of the appendices.

A.1. Additional Notation

in the proofs of our main theorems, we make use of some additional notations and standard facts about average-reward MDPs. For any policy π, its gain and bias ρ^π and h^π satisfy the optimality equations $\rho^\pi = P_\pi \rho^\pi$ and $\rho^\pi + h^\pi = r_\pi + P_\pi h^\pi$, where the second equation is sometimes called the (average-reward) Bellman equation or Poisson equation. We let $P_\pi^\infty = \text{C-lim}_{T\to\infty}(P_\pi)^T$ denote the limiting matrix of the Markov chain induced by the policy π. When this Markov chain is aperiodic, the Cesaro limit, C-lim, can be replaced with the usual limit. Note that $P_\pi^\infty P_\pi = P_\pi P_\pi^\infty = P_\pi^\infty$ and $\rho^\pi = P_\pi^\infty r_\pi$. For any policy π we define the Bellman consistency operator $\mathcal{T}_\gamma^\pi : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ by $\mathcal{T}_\gamma^\pi(x) = r_\pi + \gamma P_\pi x$.

For any $x \in \mathbb{R}^S$ and any policy π, we define a policy-specific next-state transition variance vector $\mathbb{V}_{P_\pi}[x] \in \mathbb{R}^S$ as $(\mathbb{V}_{P_\pi}[x])_s := \sum_{s'\in\mathcal{S}} (P_\pi)_{s,s'} \left[x(s') - \sum_{s''} (P_\pi)_{s,s''} x(s'')\right]^2$. We use $\|B\|_{\infty\to\infty}$ to denote the $\|\cdot\|_\infty$ operator norm of matrix B, which is equal to the maximum ℓ^1-norms of any row of B.

A.2. Complexity Parameters

In this section, we provide definitions for the diameter D and the uniform mixing time τ_{unif}, which are standard complexity parameters (along with optimal bias span $\|h^\star\|_{\text{span}}$) used in prior work on average-reward MDPs. We also give the definitions for the quantities $\|h^{\widehat{\pi}}\|_{\text{span}}$, $\|\widehat{\underline{h}}^\star\|_{\text{span}}$, and $\|\underline{h}^{\widehat{\pi}}\|_{\text{span}}$ appearing in Table 1, which arise in the results of Neu and Okolo (2024) and Zurek and Chen (2024). The results in the present paper do not involve these parameters; their definitions are included here for completeness and for comparison with our results.

First we define the diameter. For any state $s \in \mathcal{S}$, let $\eta_s = \inf\{t \geq 1 : S_t = s\}$ denote the hitting time of state s, which is a random variable (in the probability space of trajectories in the MDP P). Then the diameter D is defined as

$$D := \max_{s_1 \neq s_2} \inf_{\pi\in\Pi_{\text{MD}}} \mathbb{E}_{s_1}^\pi [\eta_{s_2}],$$

where Π_{MD} is the set of all Markovian deterministic policies. $D < \infty$ if and only if the MDP is communicating, so in particular it is generally infinite in weakly communicating MDPs.

Now we define the uniform mixing time τ_{unif}. This parameter is only defined if we first assume that for all $\pi \in \Pi_{\text{MD}}$, the Markov chain induced by P_π has a unique stationary distribution, which we call ν_π (considered as a row vector in $\mathbb{R}^{\mathcal{S}}$). Then for any policy $\pi \in \Pi_{\text{MD}}$, we can define its mixing time $\tau_\pi := \inf\left\{t \geq 1 : \max_{s\in\mathcal{S}} \left\|e_s^\top (P_\pi)^t - \nu_\pi^\top\right\|_1 \leq \frac{1}{2}\right\}$. Finally, we define the uniform mixing time $\tau_{\text{unif}} := \sup_{\pi\in\Pi_{\text{MD}}} \tau_\pi$.

It always holds that $\|h^\star\|_{\text{span}} \leq D$ (Bartlett and Tewari, 2009) and $\|h^\star\|_{\text{span}} \leq 3\tau_{\text{unif}}$ (Wang et al., 2022; Zurek and Chen, 2024). In general, D and τ_{unif} are not comparable (Wang et al., 2022), and they both can be arbitrarily larger than $\|h^\star\|_{\text{span}}$.

The prior work listed in Table 1 also involves the quantities $\|h^{\widehat{\pi}}\|_{\text{span}}$, $\|\widehat{\underline{h}}^\star\|_{\text{span}}$ and $\|\underline{h}^{\widehat{\pi}}\|_{\text{span}}$. These are all dependent on the outputs of certain (randomized) algorithms which introduce them, and

they are generally not controlled in terms of $\|h^\star\|_{\text{span}}$. In particular, the algorithm described in (Neu and Okolo, 2024, Theorem 4.1) returns a policy $\widehat{\pi}$, and their complexity bounds depend on $\|h^{\widehat{\pi}}\|_{\text{span}}$, the span of the bias of $\widehat{\pi}$ under the true MDP (P, r). Zurek and Chen (2024) define an *anchored* empirical AMDP $(\underline{\widehat{P}}, r)$ with the transition matrix $\underline{\widehat{P}} := (1-\eta)\widehat{P} + \eta \mathbf{1} e_{s_0}^\top$, where $\eta \in [0,1]$ is an algorithmic parameter, s_0 is an arbitrary state and $e_{s_0} \in \mathbb{R}^S$ is the vector with all zeros except for the entry s_0 being equal to 1. Letting $\underline{\widehat{h}}^\star$ denote the optimal bias function in this anchored AMDP, and letting $\underline{h}^{\widehat{\pi}}$ denote the bias function of the policy $\widehat{\pi}$ output by (the anchored and unperturbed version of) Algorithm 1 in Zurek and Chen (2024), their sample complexity guarantees are given in terms of the quantities $\|\underline{\widehat{h}}^\star\|_{\text{span}}$ and $\|\underline{h}^{\widehat{\pi}}\|_{\text{span}}$. Note that Zurek and Chen (2024, Theorem 14) gives an example where both of these terms are much larger than $\|h^\star\|_{\text{span}}$.

We refer to Neu and Okolo (2024) and Zurek and Chen (2024) for the complete definitions of $\|h^{\widehat{\pi}}\|_{\text{span}}$, $\|\underline{\widehat{h}}^\star\|_{\text{span}}$ and $\|\underline{h}^{\widehat{\pi}}\|_{\text{span}}$ as well as their relationship with other complexity parameters.

A.3. Fixed-n vs. Fixed-ε Settings

Here we compare the fixed-n and fixed-ε settings. Suppose we have some algorithm for the fixed-n setting which outputs a policy $\widehat{\pi}$. If the complexity parameters appearing in the suboptimality guarantee for this algorithm (e.g., $C\sqrt{\|h^\star\|_{\text{span}} \log(1/\delta)/n}$) are known, then the algorithm can be converted to the fixed-ε setting by "inverting" this bound to find n large enough to make it $\leq \varepsilon$. For the minimax rate of solving discounted MDPs of $\widetilde{O}(\sqrt{\frac{1}{(1-\gamma)^3 n}})$ (Azar et al., 2013), since γ is known, this is always possible. However, this is not generally possible in the average-reward setting, since the bounds involve unknown complexity parameters such as the optimal bias span $\|h^\star\|_{\text{span}}$, the uniform mixing time τ_{unif}, or the diameter D. A natural remedy is to attempt to estimate these parameters and then use the estimate to invert the suboptimality bound, which is the approach taken by Tuynman et al. (2024) to obtain an algorithm for the fixed-ε setting which has a D-based sample complexity, which is possible since D is estimable. However, the optimal bias span is not estimable (Tuynman et al., 2024; Zurek and Chen, 2025) and the uniform mixing time does not seem to be efficiently computable, so this approach does not seem to work. But even if the "inversion" cannot be performed algorithmically, we can still use it from an analytical perspective to compare algorithms, as we do in Table 1. We note then that some algorithms from Table 1 that are prior-knowledge-free for the fixed-n setting cannot be converted to the fixed-ε setting without complexity parameter knowledge. (The goal of removing prior knowledge is still nontrivial for the fixed-n setting; the algorithm of Zurek and Chen (2025) requires prior knowledge in the fixed-n setting in order to tune the discount factor used within a discounted reduction.) Finally, we briefly note that while lower bounds for average-reward RL have been developed for the fixed-ε setting (e.g., Wang et al. 2022), it is straightforward to convert them to the fixed-n setting.

A.4. Remark on α

We discuss the origin of the function α that appears in our algorithms and bounds.

Remark 1 *The quantity* $\alpha(\delta, n) = 96\sqrt{\log(24SAn^5/\delta)}\log_2(\log_2(n+4))$ *appearing in our bounds arises from our choice to make use of concentration bounds developed by Zurek and Chen (2024). This quantity is used within our algorithms, so we make it explicit but did not attempt to optimize it. We note Zurek and Chen (2024) also did not optimize constants/log factors, and thus we*

conjecture that a smaller function α may be sufficient for their inequalities to hold, in which case the improvement could be carried over to our work in a black-box manner.

A.5. Guide to Appendices

In Appendix B, we discuss the relationship between the $\widehat{\gamma}$ selected by our algorithm and the true optimal bias span $\|h^\star\|_{\text{span}}$. Appendix C contains the proofs for all of our main theorems, and Appendix D contains examples mentioned in Subsection 3.3 of situations where the guarantee of Theorem 4 could be much better than the minimax rate. Within Appendix C, we first prove Theorem 2 in Appendix C.1, then Theorem 1 in Appendix C.2. In Appendix C.3 we prove Lemma 3 regarding the properties of the span-constrained planning subroutine Algorithm 3, and finally in Appendix C.4 we prove Theorem 4.

Appendix B. Relationship Between $\widehat{\gamma}_i$ and $\|h^\star\|_{\text{span}}$

Recall equation (10) from the proof sketch of Theorem 2. There, we see that the discount factor $\widehat{\gamma}_i$ selected by Algorithm 2 in a given iteration i is (approximately) the minimizer of a function $\widehat{B}(\gamma)$ that upper-bounds the suboptimality of a policy $\widehat{\pi}_\gamma$ in the following manner:

$$\left\|\rho^{\widehat{\pi}_\gamma} - \rho^\star\right\|_\infty \lesssim \widehat{B}(\gamma) := (1-\gamma)\|\widehat{V}_\gamma^\star\|_{\text{span}} + \sqrt{\frac{\|\widehat{V}_\gamma^\star\|_{\text{span}} + 1}{n_i}} + \frac{1}{(1-\gamma)n_i}.$$

We can understand $\widehat{B}(\gamma)$ by relating it to the following (deterministic) quantity

$$B(\gamma) := (1-\gamma)\|V_\gamma^\star\|_{\text{span}} + \sqrt{\frac{\|V_\gamma^\star\|_{\text{span}} + 1}{n_i}} + \frac{1}{(1-\gamma)n_i}. \tag{12}$$

Using an error bound of the form $\left\|V_\gamma^\star - \widehat{V}_\gamma^\star\right\|_\infty \lesssim \frac{1}{1-\gamma}\sqrt{\frac{\|V_\gamma^\star\|_{\text{span}} + \|\widehat{V}_\gamma^\star\|_{\text{span}} + 1}{n_i}}$ (Zurek and Chen, 2024, Theorem 9), as well as the "localization" bound $\|V_\gamma^\star\|_{\text{span}} \lesssim \|\widehat{V}_\gamma^\star\|_{\text{span}} + \frac{1}{(1-\gamma)^2 n_i}$ which appears in the proof sketch, we can show that (with high probability) $B(\gamma) \lesssim \widehat{B}(\gamma)$, and likewise repeating the same steps but with the roles of $\widehat{V}_\gamma^\star$ and $V_\gamma^\star$ reversed we can also show that $\widehat{B}(\gamma) \lesssim B(\gamma)$. Therefore, since $\widehat{B}(\gamma)$ and $B(\gamma)$ are equivalent up to logarithmic factors with high probability, we can understand $\widehat{\gamma}_i$ by considering the γ which minimizes $B(\gamma)$.

We now turn to the "oracle choice" of $\gamma^\star$ made in Zurek and Chen (2024) (also discussed in our proof sketches) such that $\frac{1}{1-\gamma^\star} \approx \sqrt{n_i(\|h^\star\|_{\text{span}} + 1)}$. Using the fact that $\|V_\gamma^\star\|_{\text{span}} \lesssim \|h^\star\|_{\text{span}}$ (Wei et al., 2020), we are guaranteed the final term in the definition (12) of $B(\gamma^\star)$ is the largest and equal to $\sqrt{\frac{\|h^\star\|_{\text{span}}+1}{n_i}}$. This implies that the $\widehat{\gamma}_i$ selected by minimizing $\widehat{B}(\gamma)$ (and also the minimizing γ of $B(\gamma)$) must satisfy $\frac{1}{1-\widehat{\gamma}_i} \lesssim \frac{1}{1-\gamma^\star} \approx \sqrt{n_i(\|h^\star\|_{\text{span}} + 1)}$, since for all $\gamma \geq \gamma^\star$ the final term of $B(\gamma)$ is $\gtrsim \frac{1}{(1-\gamma^\star)n_i} \gtrsim B(\gamma^\star)$. However, $\frac{1}{1-\widehat{\gamma}_i}$ may potentially be much smaller than $\sqrt{n_i(\|h^\star\|_{\text{span}} + 1)}$, so in particular it is not generally possible to convert $\widehat{\gamma}_i$ into some constant-factor estimate of $\|h^\star\|_{\text{span}}$ (which would contradict the hardness of estimating $\|h^\star\|_{\text{span}}$ as established in Tuynman et al. 2024; Zurek and Chen 2025). Indeed, it is possible to compute the minimizer $\tilde{\gamma}^\star$ of $B(\gamma)$ for the MDP

instances that were used to show the hardness of estimating $\|h^\star\|_{\text{span}}$ in Zurek and Chen (2025, Proof of Theorem 3). We find that such instances (which have n_i as an input parameter) have $\left\|V_\gamma^\star\right\|_{\text{span}} \lesssim 1 + \frac{1}{1-\gamma}\frac{1}{\sqrt{n_i}}$, and for relatively small γ this bound can be much better than the bound $\left\|V_\gamma^\star\right\|_{\text{span}} \lesssim \|h^\star\|_{\text{span}}$; consequently, our sharper DMDP reduction (14), which replaces $\|h^\star\|_{\text{span}}$ with $\left\|V_\gamma^\star\right\|_{\text{span}}$, gives a much better bound (also see (9) in the proof sketches). This fact about $\left\|V_\gamma^\star\right\|_{\text{span}}$ for these instances also implies that the minimizing $\tilde{\gamma}^\star$ has $\frac{1}{1-\tilde{\gamma}^\star} \approx \sqrt{n_i} \ll \sqrt{n_i(\|h^\star\|_{\text{span}} + 1)}$. Since one of the instances used in Zurek and Chen (2025, Proof of Theorem 3) has an arbitrarily large bias span $\|h^\star\|_{\text{span}}$, this in particular implies that $\frac{1}{1-\tilde{\gamma}^\star}$ (and thus also $\frac{1}{1-\widehat{\gamma}_i}$) may be arbitrarily smaller than $\sqrt{n_i(\|h^\star\|_{\text{span}} + 1)}$. This minimizing choice $\tilde{\gamma}^\star$ gives $B(\tilde{\gamma}^\star) \lesssim \frac{1}{\sqrt{n_i}} \ll \sqrt{\frac{\|h^\star\|_{\text{span}}}{n_i}}$, which implies that these instances, while being worst-case for estimating the optimal bias span, are *not* worst-case for the task of finding a gain-optimal policy or estimating the optimal gain.

In summary, although $\widehat{\gamma}_i$ is chosen in our algorithm to calibrate certain approximation and estimation error terms at least as well as the "oracle choice" $\gamma^\star$ (which depends on $\|h^\star\|_{\text{span}}$), this $\widehat{\gamma}_i$ should not be seen as estimating $\|h^\star\|_{\text{span}}$, but rather as calibrating tighter bounds on approximation/estimation errors; in particular, as discussed in the proof sketches, these tighter bounds replace $\|h^\star\|_{\text{span}}$ with $\left\|V_\gamma^\star\right\|_{\text{span}}$. In fact, the "oracle choice" $\frac{1}{1-\gamma^\star} \approx \sqrt{n_i(\|h^\star\|_{\text{span}} + 1)}$ is not only more difficult to compute, but can generally achieve a worse tradeoff than the $\tilde{\gamma}^\star$ which minimizes $B(\gamma)$. Still, for worst-case instances such as those appearing in the minimax lower bounds for learning a gain-optimal policy or estimating the optimal gain, the oracle $\gamma^\star$ must indeed approximately minimize $B(\gamma)$, since we know for such instances that $B(\gamma) \gtrsim \sqrt{\|h^\star\|_{\text{span}}/n_i}$ for all γ (since they are hard instances), and also that $B(\gamma^\star) \lesssim \sqrt{\|h^\star\|_{\text{span}}/n_i}$. This observation suggests that for instances that are hard for optimizing/estimating the gain, we may actually be able to estimate $\|h^\star\|_{\text{span}}$ to a constant factor using $\widehat{\gamma}_i$, which would imply such instances are *not* hard for the task of span estimation.

Appendix C. Proofs of Main Theorems

In this section, we prove our main theorems.

C.1. Proof of Theorem 2

The proof has two main steps. First we will show that all upper and lower bounds computed within Algorithm 2 are valid upper/lower bounds on $\rho^\star$ and the gain of the output policy. This implies that whenever the algorithm terminates, it will have output a policy which is ε-optimal. The second step is to show that once a sufficient number of iterations have occurred (and thus a sufficiently large sample size is chosen), the confidence interval corresponding to some discount factor $\gamma^\star$ will be sufficiently small, thus ensuring termination of the algorithm at or before this point.

The following lemma is very similar to results of Wei et al. (2020) and Wang et al. (2022). While the bound $\left\|\rho^\star - (1-\gamma)V_\gamma^\star\right\|_\infty \le (1-\gamma)\|h^\star\|_{\text{span}}$ has appeared previously (Wei et al., 2020, Lemma 2), to the best of our knowledge a bound of the form $\left\|\rho^\star - (1-\gamma)V_\gamma^\star\right\|_\infty \le (1-\gamma)\left\|V_\gamma^\star\right\|_{\text{span}}$ has not previously appeared in the literature. As explained in the proof sketches in Section 4, the difference is extremely significant in our situation, since $\left\|V_\gamma^\star\right\|_{\text{span}}$ can be estimated while $\|h^\star\|_{\text{span}}$

generally cannot be (Zurek and Chen, 2025; Tuynman et al., 2024). This new result thus provides a computable upper bound for the "approximation error" due to reducing the AMDP to a DMDP with a certain discount factor, which can be used algorithmically. We also note that (16) below is already known and appears in (Wang et al., 2022, Lemma 6). (Their proof has a minor issue because it is not the case that all policies π have ρ^π which is state-independent/constant in a weakly communicating MDP, but their proof does not actually need this fact and the result is still true.)

Lemma 5 *Fix a discount factor $\gamma \in [0, 1)$. Then*

$$(1-\gamma)\left(\min_s V_\gamma^\star(s)\right)\mathbf{1} \le \rho^\star \le (1-\gamma)\left(\max_s V_\gamma^\star(s)\right)\mathbf{1} \tag{13}$$

which implies

$$\left\|\rho^\star - (1-\gamma)V_\gamma^\star\right\|_\infty \le (1-\gamma)\left\|V_\gamma^\star\right\|_{\text{span}}. \tag{14}$$

Additionally, for any fixed policy π, we have that

$$(1-\gamma)\left(\min_s V_\gamma^\pi(s)\right)\mathbf{1} \le \rho^\pi \le (1-\gamma)\left(\max_s V_\gamma^\pi(s)\right)\mathbf{1}, \tag{15}$$

which implies

$$\left\|\rho^\pi - (1-\gamma)V_\gamma^\pi\right\|_\infty \le (1-\gamma)\left\|V_\gamma^\pi\right\|_{\text{span}}. \tag{16}$$

Consequently, for any policy π, we have that

$$\rho^\pi \ge \rho^\star - (1-\gamma)\left(\left\|V_\gamma^\pi - V_\gamma^\star\right\|_\infty + \left\|V_\gamma^\star\right\|_{\text{span}}\right)\mathbf{1}.$$

Proof First, for the statement with a fixed policy π, as shown in Wang et al. (2022) we have that

$$\rho^\pi = P_\pi^\infty r_\pi = (1-\gamma)\sum_{t=0}^\infty \gamma^t P_\pi^\infty r_\pi = (1-\gamma)P_\pi^\infty \sum_{t=0}^\infty \gamma^t P_\pi^t r_\pi = (1-\gamma)P_\pi^\infty V_\gamma^\pi$$

which implies that, for each $s \in \mathcal{S}$, $\rho^\pi(s) = e_s^\top P_\pi^\infty \left((1-\gamma)V_\gamma^\pi\right)$, and thus $\rho^\pi(s)$ is a convex combination of entries of $(1-\gamma)V_\gamma^\pi$, implying that $(1-\gamma)\min_{s'} V_\gamma^\pi(s') \le \rho^\pi(s) \le (1-\gamma)\max_{s'} V_\gamma^\pi(s')$. By subtracting $(1-\gamma)V_\gamma^\pi(s)$ we obtain that

$$(1-\gamma)\left(\min_{s'} V_\gamma^\pi(s') - V_\gamma^\pi(s)\right) \le \rho^\pi(s) - (1-\gamma)V_\gamma^\pi(s) \le (1-\gamma)\left(\max_{s'} V_\gamma^\pi(s') - V_\gamma^\pi(s)\right)$$

which can be further lower and upper bounded as

$$-(1-\gamma)\left\|V_\gamma^\pi\right\|_{\text{span}} \le \rho^\pi(s) - (1-\gamma)V_\gamma^\pi(s) \le (1-\gamma)\left\|V_\gamma^\pi\right\|_{\text{span}}$$

implying (16).

Now we show the statement (13). For any $s \in \mathcal{S}$, we similarly have

$$\rho^\star(s) = (1-\gamma)e_s^\top P_{\pi^\star}^\infty V_\gamma^{\pi^\star} \le (1-\gamma)e_s^\top P_{\pi^\star}^\infty V_\gamma^\star \le (1-\gamma)\max_s V_\gamma^\star(s)$$

and also for any $s \in \mathcal{S}$

$$\rho^\star(s) \geq \rho^{\pi^\star_\gamma}(s) = (1-\gamma) e_s^\top P^\infty_{\pi^\star_\gamma} V^\star_\gamma \geq (1-\gamma) \min_s V^\star_\gamma(s).$$

Statement (14) follows in an identical manner as to how (16) followed from (15).

For the final desired statement, using the previous results we have

$$\begin{aligned}
\rho^\pi &\geq (1-\gamma) \min_s V^\pi_\gamma(s) \mathbf{1} \geq (1-\gamma) \min_s V^\star_\gamma(s) \mathbf{1} - (1-\gamma) \left\| V^\pi_\gamma - V^\star_\gamma \right\|_\infty \mathbf{1} \\
&\geq (1-\gamma) \max_s V^\star_\gamma(s) \mathbf{1} - (1-\gamma) \left\| V^\star_\gamma \right\|_{\text{span}} \mathbf{1} - (1-\gamma) \left\| V^\pi_\gamma - V^\star_\gamma \right\|_\infty \mathbf{1} \\
&\geq \rho^\star - (1-\gamma) \left\| V^\star_\gamma \right\|_{\text{span}} \mathbf{1} - (1-\gamma) \left\| V^\pi_\gamma - V^\star_\gamma \right\|_\infty \mathbf{1}.
\end{aligned}$$

■

Here we repeat the definitions of some quantities which are defined in Algorithm 2, and additionally we extend their definitions to hold for all integers $i \geq 1$ (rather than just the iterations which appear in the course of the algorithm before its termination). For all $i \geq 1$, we define $\widehat{P}^{(i)}$ to be constructed in the manner shown in Algorithm 2 via taking $n_i = 2^i$ samples from all state-action pairs. We also let $\mathcal{H}_i := \{\gamma : \text{there exists an integer } k \text{ such that } \sqrt{n_i} \leq \frac{1}{1-\gamma} = 2^k \leq n_i\}$. We can then define, for all $\gamma \in \mathcal{H}_i$, the policy $\widetilde{\pi}_{\gamma,i}$ and value function $\widetilde{V}_{\gamma,i}$ as the outputs from SOLVEDMDP$(\widehat{P}^{(i)}, r, \gamma, \frac{1}{n_i})$. With such quantities we can then define $\widehat{U}_i(\gamma), \widehat{L}_i(\gamma)$ following the definitions given in 2, and finally we let $\widehat{\gamma}_i := \text{argmin}_{\gamma \in \mathcal{H}_i} \widehat{U}_i(\gamma) - \widehat{L}_i(\gamma)$. We also use $\widehat{V}^\pi_{\gamma,i}$ to denote the value function of policy π in the DMDP $(\widehat{P}^{(i)}, r, \gamma)$ and $\widehat{V}^\star_{\gamma,i}$ to denote the optimal value function of this DMDP.

Lemma 6 *Define the function* $\alpha(\tilde{\delta}, \tilde{n}) = 24\sqrt{16 \log\left(\frac{24SA\tilde{n}^5}{\tilde{\delta}}\right)} \log_2(\log_2(\tilde{n}+4))$. *Fix* $\delta > 0$. *Then with probability at least* $1-\delta$, *we have for all integers* $i \geq 1$ *and all* $\gamma \in \mathcal{H}_i$ *that*

$$\left\| V^{\pi^\star_\gamma}_\gamma - \widehat{V}^{\pi^\star_\gamma}_{\gamma,i} \right\|_\infty \leq \frac{\alpha(\delta, n_i)}{1-\gamma} \sqrt{\frac{\left\| V^{\pi^\star_\gamma}_\gamma \right\|_{\text{span}} + 1}{n_i}} \tag{17}$$

and also the subroutine on line 10 outputs a policy $\widetilde{\pi}_{\gamma,i}$ *such that*

$$\widehat{V}^{\widetilde{\pi}_{\gamma,i}}_{\gamma,i} \geq \widehat{V}^\star_{\gamma,i} - \frac{1}{n_i}\mathbf{1} \tag{18}$$

$$\left\| \widetilde{V}_{\gamma,i} - \widehat{V}^\star_{\gamma,i} \right\|_\infty \leq \frac{1}{n_i} \tag{19}$$

$$\left\| \widehat{V}^{\widetilde{\pi}_{\gamma,i}}_{\gamma,i} - V^{\widetilde{\pi}_{\gamma,i}}_\gamma \right\|_\infty \leq \frac{\alpha(\delta, n_i)}{1-\gamma} \sqrt{\frac{\left\| \widehat{V}^{\widetilde{\pi}_{\gamma,i}}_{\gamma,i} \right\|_{\text{span}} + 1}{n_i}}. \tag{20}$$

Proof We fix a pair i and $\gamma \in \mathcal{H}_i$ and establish the desired bounds, and then we will take a union bound. From (Zurek and Chen, 2024, Proof of Theorem 9), in particular (Zurek and Chen, 2024, Equations (31) and (32)), we have with probability at least $1-2\delta$ that

$$\left\| V^{\pi^\star_\gamma}_\gamma - \widehat{V}^{\pi^\star_\gamma}_\gamma \right\|_\infty \leq \frac{24 \log_2 \log_2\left(\frac{1}{1-\gamma} + 4\right)}{1-\gamma} \sqrt{16 \frac{\left\| V^\star_\gamma \right\|_{\text{span}} + 1}{n_i} \log\left(\frac{12SAn_i}{(1-\gamma)^2\delta}\right)}$$

and

$$\left\|V_\gamma^{\widetilde{\pi}_\gamma} - \widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_\infty \leq \frac{24\log_2\log_2\left(\frac{1}{1-\gamma}+4\right)}{1-\gamma}\sqrt{16\frac{\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}}+1}{n_i}\log\left(\frac{12SAn_i}{(1-\gamma)^2\delta}\right)}.$$

By definition of $\mathcal{H}_i$ we have that $\frac{1}{1-\gamma} \leq n_i$, which we use to simplify the bounds slightly to

$$\left\|V_\gamma^{\pi_\gamma^\star} - \widehat{V}_\gamma^{\pi_\gamma^\star}\right\|_\infty \leq \frac{24\log_2\log_2\left(n_i+4\right)}{1-\gamma}\sqrt{16\frac{\left\|V_\gamma^\star\right\|_{\text{span}}+1}{n_i}\log\left(\frac{12SAn_i^3}{\delta}\right)} \tag{21}$$

and

$$\left\|V_\gamma^{\widetilde{\pi}_\gamma} - \widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_\infty \leq \frac{24\log_2\log_2\left(n_i+4\right)}{1-\gamma}\sqrt{16\frac{\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}}+1}{n_i}\log\left(\frac{12SAn_i^3}{\delta}\right)}. \tag{22}$$

Now keeping i fixed and taking a union bound over all $\gamma \in \mathcal{H}_i$, we have that inequalities (21) and (22) hold for all $\gamma \in \mathcal{H}_i$ with probability at least $1 - 2\delta|\mathcal{H}_i|$. Also

$$|\mathcal{H}_i| \leq 1 + \log_2\frac{n_i}{\sqrt{n_i}} \leq 1 + \frac{1}{2}\log_2 n_i \leq 1 + \frac{1}{2}n_i \leq n_i$$

using the facts that $\log_2 x \leq x$ and that $n_i \geq 2$ (so $1 \leq \frac{1}{2}n_i$). Now adjusting the failure probability, we have (for any $\delta_i > 0$) that with probability at least $1 - \delta_i$, for all $\gamma \in \mathcal{H}_i$, both

$$\left\|V_\gamma^{\pi_\gamma^\star} - \widehat{V}_\gamma^{\pi_\gamma^\star}\right\|_\infty \leq \frac{24\log_2\log_2\left(n_i+4\right)}{1-\gamma}\sqrt{16\frac{\left\|V_\gamma^\star\right\|_{\text{span}}+1}{n_i}\log\left(\frac{24SAn_i^4}{\delta_i}\right)} \tag{23}$$

and

$$\left\|V_\gamma^{\widetilde{\pi}_\gamma} - \widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_\infty \leq \frac{24\log_2\log_2\left(n_i+4\right)}{1-\gamma}\sqrt{16\frac{\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}}+1}{n_i}\log\left(\frac{24SAn_i^4}{\delta_i}\right)} \tag{24}$$

are true. Now we union bound over all natural numbers $i \geq 1$. We set $\delta_i = \frac{\delta}{2^i} = \frac{\delta}{n_i}$ and thus obtain that with probability at least $1 - \sum_{i\geq 1}\delta_i = 1 - \delta$, for all $i \geq 1$ and all $\gamma \in \mathcal{H}_i$ we have both (23) and (24). By our choice of δ_i's these can be written as

$$\begin{aligned}\left\|V_\gamma^{\pi_\gamma^\star} - \widehat{V}_\gamma^{\pi_\gamma^\star}\right\|_\infty &\leq \frac{24\log_2\log_2\left(n_i+4\right)}{1-\gamma}\sqrt{16\frac{\left\|V_\gamma^\star\right\|_{\text{span}}+1}{n_i}\log\left(\frac{24SAn_i^5}{\delta}\right)} \\ &= \frac{\alpha(\delta,n_i)}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}}+1}{n_i}}\end{aligned}$$

and

$$\begin{aligned}\left\|V_\gamma^{\widetilde{\pi}_\gamma} - \widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_\infty &\leq \frac{24\log_2\log_2\left(n_i+4\right)}{1-\gamma}\sqrt{16\frac{\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}}+1}{n_i}\log\left(\frac{24SAn_i^5}{\delta}\right)} \\ &= \frac{\alpha(\delta,n_i)}{1-\gamma}\sqrt{\frac{\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}}+1}{n_i}}\end{aligned}$$

as desired. ■

Lemma 7 *Suppose that there exists some $m, \beta > 1$, $\gamma < 1$, policy π, MDPs (P, r), $(\overline{P}, r)$, such that*

$$\left\|V_\gamma^{\pi_\gamma^\star} - \overline{V}_\gamma^{\pi_\gamma^\star}\right\|_\infty \leq \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^{\pi_\gamma^\star}\right\|_{\text{span}} + 1}{m}} \tag{25}$$

$$\overline{V}_\gamma^{\pi} \geq \overline{V}_\gamma^{\star} - \frac{1}{m}\mathbf{1} \tag{26}$$

$$\left\|\overline{V}_\gamma^{\pi} - V_\gamma^{\pi}\right\|_\infty \leq \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|\overline{V}_\gamma^{\pi}\right\|_{\text{span}} + 1}{m}} \tag{27}$$

where for any π' we use $\overline{V}_\gamma^{\pi'}$ to denote the value of policy π' in the DMDP $(\overline{P}, r, \gamma)$, where $\overline{V}_\gamma^\star$ denotes the optimal value function in the DMDP $(\overline{P}, r, \gamma)$, and where $\pi_\gamma^\star$ is the optimal policy for the DMDP (P, r, γ). Then

$$\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty \leq \frac{2\beta^2}{(1-\gamma)^2 m} + \frac{4\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 1 + \frac{1}{m}}{m}} + \frac{4}{m} \tag{28}$$

and

$$\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty \leq \frac{2\beta^2}{(1-\gamma)^2 m} + \frac{4\beta}{1-\gamma}\sqrt{\frac{\left\|\overline{V}_\gamma^\star\right\|_{\text{span}} + 1 + \frac{1}{m}}{m}} + \frac{4}{m}. \tag{29}$$

Proof Using the triangle inequality several times and (26), we have

$$\begin{aligned}
V_\gamma^\star &\geq V_\gamma^\pi \\
&\geq \overline{V}_\gamma^\pi - \left\|\overline{V}_\gamma^\pi - V_\gamma^\pi\right\|_\infty \mathbf{1} \\
&\geq \overline{V}_\gamma^\star - \frac{1}{m}\mathbf{1} - \left\|\overline{V}_\gamma^\pi - V_\gamma^\pi\right\|_\infty \mathbf{1} \\
&\geq \overline{V}_\gamma^{\pi_\gamma^\star} - \frac{1}{m}\mathbf{1} - \left\|\overline{V}_\gamma^\pi - V_\gamma^\pi\right\|_\infty \mathbf{1} \\
&\geq V_\gamma^{\pi_\gamma^\star} - \left\|V_\gamma^{\pi_\gamma^\star} - \overline{V}_\gamma^{\pi_\gamma^\star}\right\|_\infty \mathbf{1} - \frac{1}{m}\mathbf{1} - \left\|\overline{V}_\gamma^\pi - V_\gamma^\pi\right\|_\infty \mathbf{1}.
\end{aligned}$$

Since by definition $V_\gamma^{\pi_\gamma^\star} = V_\gamma^\star$, subtracting this term from all expressions we obtain

$$\mathbf{0} \geq \overline{V}_\gamma^\star - V_\gamma^\star - \frac{1}{m}\mathbf{1} - \left\|\overline{V}_\gamma^\pi - V_\gamma^\pi\right\|_\infty \mathbf{1} \geq -\left\|V_\gamma^{\pi_\gamma^\star} - \overline{V}_\gamma^{\pi_\gamma^\star}\right\|_\infty \mathbf{1} - \frac{1}{m}\mathbf{1} - \left\|\overline{V}_\gamma^\pi - V_\gamma^\pi\right\|_\infty \mathbf{1}$$

which after rearranging implies that

$$\frac{1}{m}\mathbf{1} + \left\|\overline{V}_\gamma^\pi - V_\gamma^\pi\right\|_\infty \mathbf{1} \geq \overline{V}_\gamma^\star - V_\gamma^\star \geq -\left\|V_\gamma^{\pi_\gamma^\star} - \overline{V}_\gamma^{\pi_\gamma^\star}\right\|_\infty \mathbf{1}$$

which further implies

$$\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty \le \left\|\overline{V}_\gamma^\pi - V_\gamma^\pi\right\|_\infty + \left\|V_\gamma^{\pi_\gamma^\star} - \overline{V}_\gamma^{\pi_\gamma^\star}\right\|_\infty + \frac{1}{m}. \tag{30}$$

Now we will focus on establishing the first inequality (28) in the lemma statement. For the first term on the RHS of (30), we can use condition (27) and then triangle inequality to bound

$$\begin{aligned}\left\|\overline{V}_\gamma^\pi - V_\gamma^\pi\right\|_\infty &\le \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|\overline{V}_\gamma^\pi\right\|_{\text{span}} + 1}{m}} \\ &\le \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + \left\|\overline{V}_\gamma^\pi - V_\gamma^\star\right\|_{\text{span}} + 1}{m}} \\ &\le \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 2\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty + \frac{1}{m} + 1}{m}}\end{aligned} \tag{31}$$

using that

$$\left\|\overline{V}_\gamma^\pi - V_\gamma^\star\right\|_{\text{span}} \le \left\|\overline{V}_\gamma^\pi - \overline{V}_\gamma^\star\right\|_{\text{span}} + \left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_{\text{span}} \le \frac{1}{m} + 2\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty$$

since $\|\cdot\|_{\text{span}} \le 2\|\cdot\|_\infty$ and $\mathbf{0} \le \overline{V}_\gamma^\star - \overline{V}_\gamma^\pi \le \frac{1}{m}\mathbf{1}$. For the second term on the RHS of (30), we can use the condition (25). Thus combining (30), (31), and (25), we obtain

$$\begin{aligned}\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty &\le \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 2\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty + 1 + \frac{1}{m}}{m}} + \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^{\pi_\gamma^\star}\right\|_{\text{span}} + 1}{m}} + \frac{1}{m} \\ &\le \frac{2\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 1 + \frac{1}{m}}{m}} + \frac{\beta}{1-\gamma}\sqrt{\frac{2\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty}{m}} + \frac{1}{m}\end{aligned} \tag{32}$$

where we used that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ (and $V_\gamma^\star = V_\gamma^{\pi_\gamma^\star}$). Rearranging (32) as

$$\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty - \frac{\sqrt{2}\beta}{(1-\gamma)\sqrt{m}}\sqrt{\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty} - \frac{2\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 1 + \frac{1}{m}}{m}} + \frac{1}{m} \le 0$$

and then using the quadratic formula to find the largest possible root of this quadratic polynomial in $\sqrt{\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty}$, we obtain that

$$\sqrt{\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty} \le \frac{1}{2}\frac{\sqrt{2}\beta}{(1-\gamma)\sqrt{m}} + \frac{1}{2}\sqrt{\left(\frac{\sqrt{2}\beta}{(1-\gamma)\sqrt{m}}\right)^2 + 4\frac{2\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 1 + \frac{1}{m}}{m}} + \frac{4}{m}}. \tag{33}$$

Now squaring both sides and using that $(a+b)^2 \leq 2a^2 + 2b^2$ which implies $(\frac{a}{2} + \frac{b}{2})^2 \leq \frac{a^2}{2} + \frac{b^2}{2}$, we obtain

$$\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty \leq \frac{1}{2}\frac{2\beta^2}{(1-\gamma)^2 m} + \frac{1}{2}\left(\left(\frac{\sqrt{2}\beta}{(1-\gamma)\sqrt{m}}\right)^2 + 4\frac{2\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 1 + \frac{1}{m}}{m}} + \frac{4}{m}\right)$$
$$= \frac{2\beta^2}{(1-\gamma)^2 m} + \frac{4\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 1 + \frac{1}{m}}{m}} + \frac{4}{m} \tag{34}$$

as desired.

We next focus on establishing the second inequality (29) from the lemma statement. We now will bound the first term on the RHS of (30) in terms of $\left\|\overline{V}_\gamma^\star\right\|_{\text{span}}$ with an analogous argument but instead using only that $\left\|\overline{V}_\gamma^\pi\right\|_{\text{span}} \leq \left\|\overline{V}_\gamma^\star\right\|_{\text{span}} + \left\|\overline{V}_\gamma^\pi - \overline{V}_\gamma^\star\right\|_{\text{span}} \leq \left\|\overline{V}_\gamma^\star\right\|_{\text{span}} + \frac{1}{m}$, to obtain

$$\left\|\overline{V}_\gamma^\pi - V_\gamma^\pi\right\|_\infty \leq \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|\overline{V}_\gamma^\star\right\|_{\text{span}} + \frac{1}{m} + 1}{m}}. \tag{35}$$

We bound the second term on the RHS of (30) in terms of $\left\|\overline{V}_\gamma^\star\right\|_{\text{span}}$ by using (25) and then that $\left\|V_\gamma^{\pi_\gamma^\star}\right\|_{\text{span}} \leq \left\|\overline{V}_\gamma^\star\right\|_{\text{span}} + \left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_{\text{span}} \leq \left\|\overline{V}_\gamma^\star\right\|_{\text{span}} + 2\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty$ to obtain

$$\left\|V_\gamma^{\pi_\gamma^\star} - \overline{V}_\gamma^{\pi_\gamma^\star}\right\|_\infty \leq \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^{\pi_\gamma^\star}\right\|_{\text{span}} + 1}{m}} \leq \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|\overline{V}_\gamma^\star\right\|_{\text{span}} + 2\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty + 1}{m}}. \tag{36}$$

Combining (30), (35), and (36), we obtain

$$\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty \leq \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|\overline{V}_\gamma^\star\right\|_{\text{span}} + \frac{1}{m} + 1}{m}} + \frac{\beta}{1-\gamma}\sqrt{\frac{\left\|\overline{V}_\gamma^\star\right\|_{\text{span}} + 2\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty + 1}{m}} + \frac{1}{m}$$
$$\leq \frac{2\beta}{1-\gamma}\sqrt{\frac{\left\|\overline{V}_\gamma^\star\right\|_{\text{span}} + \frac{1}{m} + 1}{m}} + \frac{\beta}{1-\gamma}\sqrt{\frac{2\left\|\overline{V}_\gamma^\star - V_\gamma^\star\right\|_\infty}{m}} + \frac{1}{m} \tag{37}$$

using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for the second inequality. Since (37) is identical to (32) except for the presence of $\left\|\overline{V}_\gamma^\star\right\|_{\text{span}}$ instead of $\left\|V_\gamma^\star\right\|_{\text{span}}$, we can take completely analogous steps to obtain (29). ■

The following lemma is not necessary to show the validity of the lower bounds, but it is necessary for the upper bounds.

Lemma 8 *Under the event in Lemma 6, for all integers $i \geq 1$ and all $\gamma \in \mathcal{H}_i$, we have*

$$\left\|\widehat{V}_{\gamma,i}^\star - V_\gamma^\star\right\|_\infty \leq \frac{2\alpha(\delta, n_i)^2}{(1-\gamma)^2 n_i} + \frac{4\alpha(\delta, n_i)}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 1 + \frac{1}{n_i}}{n_i}} + \frac{4}{n_i} \tag{38}$$

and

$$\left\|\widehat{V}^{\star}_{\gamma,i} - V^{\star}_{\gamma}\right\|_{\infty} \leq \frac{2\alpha(\delta,n_i)^2}{(1-\gamma)^2 n_i} + \frac{4\alpha(\delta,n_i)}{1-\gamma}\sqrt{\frac{\left\|\widehat{V}^{\star}_{\gamma,i}\right\|_{\text{span}} + 1 + \frac{1}{n_i}}{n_i}} + \frac{4}{n_i} \tag{39}$$

Proof This follows immediately from Lemma 7 since the event described in Lemma 6 exactly meets the conditions of Lemma 7 (by setting $m = n_i$ and $\beta = \alpha(\delta, n_i)$, for all i and $\gamma \in \mathcal{H}_i$). ■

Now we show that under the event in Lemma 6, all lower and upper bounds constructed within Algorithm 2 are valid lower/upper bounds of $\rho^{\star}$.

Lemma 9 *Under the event in Lemma 6, for all integers $i \geq 1$ and all $\gamma \in \mathcal{H}_i$, we have*

$$\widehat{L}_i(\gamma)\mathbf{1} \leq \rho^{\widetilde{\pi}_{\gamma,i}} \leq \rho^{\star} \leq \widehat{U}_i(\gamma)\mathbf{1}. \tag{40}$$

Proof Fix an arbitrary integer $i \geq 1$ and $\gamma \in \mathcal{H}_i$. By (15), optimality of $\rho^{\star}$, and (13) we have

$$\begin{aligned}(1-\gamma)\left(\min_s V_{\gamma}^{\widetilde{\pi}_{\gamma,i}}(s)\right)\mathbf{1} &\leq \rho^{\widetilde{\pi}_{\gamma,i}} \\ &\leq \rho^{\star} \\ &\leq (1-\gamma)\left(\max_s V_{\gamma}^{\star}(s)\right)\mathbf{1}.\end{aligned} \tag{41}$$

We start by lower-bounding the LHS of (41). Using (18) and (20),

$$\begin{aligned}(1-\gamma)\left(\min_s V_{\gamma}^{\widetilde{\pi}_{\gamma,i}}(s)\right)\mathbf{1} &\geq (1-\gamma)\left(\min_s \widehat{V}_{\gamma,i}^{\widetilde{\pi}_{\gamma,i}}(s)\right)\mathbf{1} - (1-\gamma)\left\|\widehat{V}_{\gamma,i}^{\widetilde{\pi}_{\gamma,i}} - V_{\gamma}^{\widetilde{\pi}_{\gamma,i}}\right\|_{\infty}\mathbf{1} \\ &\geq (1-\gamma)\left(\min_s \widehat{V}_{\gamma,i}^{\star}(s)\right)\mathbf{1} - \frac{1-\gamma}{n_i}\mathbf{1} - (1-\gamma)\left\|\widehat{V}_{\gamma,i}^{\widetilde{\pi}_{\gamma,i}} - V_{\gamma}^{\widetilde{\pi}_{\gamma,i}}\right\|_{\infty}\mathbf{1} \\ &\geq (1-\gamma)\left(\min_s \widehat{V}_{\gamma,i}^{\star}(s)\right)\mathbf{1} - \frac{1-\gamma}{n_i}\mathbf{1} - \alpha(\delta,n_i)\sqrt{\frac{\left\|\widehat{V}_{\gamma,i}^{\widetilde{\pi}_{\gamma,i}}\right\|_{\text{span}} + 1}{n_i}}\mathbf{1}.\end{aligned} \tag{42}$$

Now we replace the quantities in (42) with observable quantities in terms of $\widetilde{V}_{\gamma,i}$. Using the requirement (18) to relate $\widehat{V}_{\gamma,i}^{\widetilde{\pi}_{\gamma,i}}$ and $\widehat{V}_{\gamma,i}^{\star}$, and then (19) which bounds $\left\|\widetilde{V}_{\gamma,i} - \widehat{V}_{\gamma,i}^{\star}\right\|_{\infty}$, we have

$$\begin{aligned}\left\|\widehat{V}_{\gamma,i}^{\widetilde{\pi}_{\gamma,i}}\right\|_{\text{span}} &\leq \left\|\widehat{V}_{\gamma,i}^{\star}\right\|_{\text{span}} + \left\|\widehat{V}_{\gamma,i}^{\star} - \widehat{V}_{\gamma,i}^{\widetilde{\pi}_{\gamma,i}}\right\|_{\text{span}} \\ &\leq \left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} + \left\|\widetilde{V}_{\gamma,i} - \widehat{V}_{\gamma,i}^{\star}\right\|_{\text{span}} + \left\|\widehat{V}_{\gamma,i}^{\star} - \widehat{V}_{\gamma,i}^{\widetilde{\pi}_{\gamma,i}}\right\|_{\text{span}} \\ &\leq \left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} + 2\left\|\widetilde{V}_{\gamma,i} - \widehat{V}_{\gamma,i}^{\star}\right\|_{\infty} + \frac{1}{n_i} \\ &\leq \left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} + \frac{2}{n_i} + \frac{1}{n_i}\end{aligned} \tag{43}$$

(where $\left\|\widehat{V}^{\star}_{\gamma,i} - \widehat{V}^{\widetilde{\pi}_{\gamma,i}}_{\gamma,i}\right\|_{\text{span}} \leq \frac{1}{n_i}$ since $\mathbf{0} \leq \widehat{V}^{\star}_{\gamma,i} - \widehat{V}^{\widetilde{\pi}_{\gamma,i}}_{\gamma,i} \leq \frac{1}{n_i}\mathbf{1}$). Using this bound (43), as well as (19) again, we can further bound (42) as

$$
\begin{aligned}
&(1-\gamma)\left(\min_s \widehat{V}^{\star}_{\gamma,i}(s)\right)\mathbf{1} - \frac{1-\gamma}{n_i}\mathbf{1} - \alpha(\delta,n_i)\sqrt{\frac{\left\|\widehat{V}^{\widetilde{\pi}_{\gamma,i}}_{\gamma,i}\right\|_{\text{span}} + 1}{n_i}}\mathbf{1} \\
&\geq (1-\gamma)\left(\min_s \widetilde{V}_{\gamma,i}(s)\right)\mathbf{1} - (1-\gamma)\left\|\widetilde{V}_{\gamma,i} - \widehat{V}^{\star}_{\gamma,i}\right\|_{\infty}\mathbf{1} - \frac{1-\gamma}{n_i}\mathbf{1} - \alpha(\delta,n_i)\sqrt{\frac{\left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} + \frac{3}{n_i} + 1}{n_i}}\mathbf{1} \\
&\geq (1-\gamma)\left(\min_s \widetilde{V}_{\gamma,i}(s)\right)\mathbf{1} - 2\frac{1-\gamma}{n_i}\mathbf{1} - \alpha(\delta,n_i)\sqrt{\frac{\left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} + \frac{3}{n_i} + 1}{n_i}}\mathbf{1} \\
&= \widehat{L}_i(\gamma)\mathbf{1} \qquad (44)
\end{aligned}
$$

Now we upper-bound the RHS of (41). The inequality (39) plays a key role. Using this bound, as well as (19) to bound $\left\|\widetilde{V}_{\gamma,i} - \widehat{V}^{\star}_{\gamma,i}\right\|_{\infty}$, and the fact that $\left\|\widehat{V}^{\star}_{\gamma,i}\right\|_{\text{span}} \leq \left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} + \left\|\widehat{V}^{\star}_{\gamma,i} - \widetilde{V}_{\gamma,i}\right\|_{\text{span}} \leq \left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} + \frac{2}{n_i}$ (since $\|\cdot\|_{\text{span}} \leq 2\|\cdot\|_{\infty}$ and using (19)), we obtain

$$
\begin{aligned}
&(1-\gamma)\left(\max_s V^{\star}_{\gamma}(s)\right)\mathbf{1} \\
&\leq (1-\gamma)\left(\max_s \widehat{V}^{\star}_{\gamma,i}(s)\right)\mathbf{1} + (1-\gamma)\left\|\widehat{V}^{\star}_{\gamma,i} - V^{\star}_{\gamma}\right\|_{\infty}\mathbf{1} \\
&\leq (1-\gamma)\left(\max_s \widetilde{V}_{\gamma,i}(s)\right)\mathbf{1} + (1-\gamma)\left\|\widetilde{V}_{\gamma,i} - \widehat{V}^{\star}_{\gamma,i}\right\|_{\infty}\mathbf{1} + (1-\gamma)\left\|\widehat{V}^{\star}_{\gamma,i} - V^{\star}_{\gamma}\right\|_{\infty}\mathbf{1} \\
&\leq (1-\gamma)\left(\max_s \widetilde{V}_{\gamma,i}(s)\right)\mathbf{1} + \frac{1-\gamma}{n_i}\mathbf{1} + \frac{2\alpha(\delta,n_i)^2}{(1-\gamma)n_i}\mathbf{1} + 4\alpha(\delta,n_i)\sqrt{\frac{\left\|\widehat{V}^{\star}_{\gamma,i}\right\|_{\text{span}} + 1 + \frac{1}{n_i}}{n_i}}\mathbf{1} + (1-\gamma)\frac{4}{n_i}\mathbf{1} \\
&\leq (1-\gamma)\left(\max_s \widetilde{V}_{\gamma,i}(s)\right)\mathbf{1} + 5\frac{1-\gamma}{n_i}\mathbf{1} + \frac{2\alpha(\delta,n_i)^2}{(1-\gamma)n_i}\mathbf{1} + 4\alpha(\delta,n_i)\sqrt{\frac{\left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} + 1 + \frac{3}{n_i}}{n_i}}\mathbf{1} \\
&= \widehat{U}_i(\gamma)\mathbf{1}. \qquad (45)
\end{aligned}
$$

Combining (41), (44), and (45), and unfixing i and γ, we obtain the desired conclusion. ■

Lemma 9 implies that whenever the algorithm terminates, the resulting policy will be ε-optimal and the resulting confidence interval will be valid and of size $\leq \varepsilon$. Next we show that the algorithm will terminate by a certain iteration, by showing that on this iteration the confidence interval corresponding to some discount factor will be small.

First we define, for all integers $i \geq 1$, the discount factor $\gamma_i^{\star}$ as

$$
\gamma_i^{\star} = \inf\left\{\gamma \in \mathcal{H}_i : \frac{1}{1-\gamma} \geq \sqrt{n_i\left(\|h^{\star}\|_{\text{span}} + 1\right)}\right\}. \qquad (46)
$$

(If $\|h^{\star}\|_{\text{span}}$ is large relative to n_i then, since the largest value of $\frac{1}{1-\gamma}$ for $\gamma \in \mathcal{H}_i$ is n_i, the above set might be empty, in which case by usual convention we would have $\gamma_i^{\star} = \inf \emptyset = \infty$.)

Lemma 10 *For all integers $i \geq 1$, if*

$$\|h^\star\|_{\text{span}} + 1 \leq n_i \tag{47}$$

then $\gamma_i^\star$ is finite, and furthermore

$$\sqrt{n_i\left(\|h^\star\|_{\text{span}} + 1\right)} \leq \frac{1}{1-\gamma_i^\star} \leq 2\sqrt{n_i\left(\|h^\star\|_{\text{span}} + 1\right)}. \tag{48}$$

Proof First, by condition (47) we have that $\sqrt{n_i\left(\|h^\star\|_{\text{span}} + 1\right)} \leq n_i$, and since the largest element of $\mathcal{H}_i$ is n_i, the set in the definition (46) of $\gamma_i^\star$ will be nonempty and thus $\gamma_i^\star$ will be finite.

Furthermore since $\|h^\star\|_{\text{span}} + 1 \geq 1$, we have $\sqrt{n_i\left(\|h^\star\|_{\text{span}} + 1\right)} \geq \sqrt{n_i}$. $\sqrt{n_i}$ may not be a member of $\mathcal{H}_i$ but the smallest power of 2 which is $\geq \sqrt{n_i}$ will be a member of $\mathcal{H}_i$. Therefore $\min \mathcal{H}_i \leq \frac{1}{1-\gamma_i^\star} \leq \max \mathcal{H}_i$, so by the construction of $\mathcal{H}_i$ in line (8) of Algorithm 2, condition (48) must hold as $\mathcal{H}_i$ contains all powers of 2 within $[\min \mathcal{H}_i, \max \mathcal{H}_i]$. ■

Lemma 11 *Under the event in Lemma 6, for all integers $i \geq 1$ such that $\|h^\star\|_{\text{span}} + 1 \leq n_i$, it holds that*

$$\widehat{U}_i(\gamma_i^\star) - \widehat{L}_i(\gamma_i^\star) \leq 90\alpha(\delta, n_i)^2\sqrt{\frac{\|h^\star\|_{\text{span}} + 1}{n_i}}.$$

In particular there exist some absolute constants C_1, C_2 such that letting

$$B = \left\lceil \log_2\left(C_1 \frac{\|h^\star\|_{\text{span}} + 1}{\varepsilon^2} \log^3\left(\frac{C_2 SA(\|h^\star\|_{\text{span}} + 1)}{\delta\varepsilon}\right)\right)\right\rceil$$

we have

$$\widehat{U}_B(\gamma_B^\star) - \widehat{L}_B(\gamma_B^\star) \leq \varepsilon.$$

Proof For any integer $i \geq 1$ and any $\gamma \in \mathcal{H}_i$ we have

$$\begin{aligned}
\widehat{U}_i(\gamma) - \widehat{L}_i(\gamma) &= (1-\gamma)\left(\max_s \widetilde{V}_{\gamma,i}(s) - \min_s \widetilde{V}_{\gamma,i}(s)\right) + 7\frac{1-\gamma}{n_i} + \frac{2\alpha(\delta,n_i)^2}{(1-\gamma)n_i} \\
&\quad + 5\alpha(\delta,n_i)\sqrt{\frac{\left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} + 1 + \frac{3}{n_i}}{n_i}} \\
&= (1-\gamma)\left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} + 7\frac{1-\gamma}{n_i} + \frac{2\alpha(\delta,n_i)^2}{(1-\gamma)n_i} + 5\alpha(\delta,n_i)\sqrt{\frac{\left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} + 1 + \frac{3}{n_i}}{n_i}}.
\end{aligned} \tag{49}$$

Now we will set $\gamma = \gamma_i^\star$ and relate all terms in (49) to $\|h^\star\|_{\text{span}}$, but first we bound $\left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}}$. Under the event of Lemma 6 we have from (19), and from (38) in Lemma 8, that

$$\begin{aligned}
\left\|\widetilde{V}_{\gamma,i}\right\|_{\text{span}} &\le \left\|V_\gamma^\star\right\|_{\text{span}} + \left\|\widetilde{V}_{\gamma,i} - V_\gamma^\star\right\|_{\text{span}} \\
&\le \left\|V_\gamma^\star\right\|_{\text{span}} + 2\left\|\widetilde{V}_{\gamma,i} - V_\gamma^\star\right\|_\infty \\
&\le \left\|V_\gamma^\star\right\|_{\text{span}} + 2\left\|\widetilde{V}_{\gamma,i} - \widehat{V}_{\gamma,i}^\star\right\|_\infty + 2\left\|\widehat{V}_{\gamma,i}^\star - V_\gamma^\star\right\|_\infty \\
&\le \left\|V_\gamma^\star\right\|_{\text{span}} + 2\frac{1}{n_i} + 2\left(\frac{2\alpha(\delta,n_i)^2}{(1-\gamma)^2 n_i} + \frac{4\alpha(\delta,n_i)}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 1 + \frac{1}{n_i}}{n_i}} + \frac{4}{n_i}\right).
\end{aligned} \tag{50}$$

Bounding (50) by substituting $\gamma = \gamma_i^\star$ and using Lemma 10 to bound $\frac{1}{1-\gamma_i^\star}$, and using that $\left\|V_\gamma^\star\right\|_{\text{span}} \le 2\left\|h^\star\right\|_{\text{span}}$ (Wei et al., 2020, Lemma 2), we have

$$\begin{aligned}
\left\|\widetilde{V}_{\gamma_i^\star,i}\right\|_{\text{span}} &\le 2\left\|h^\star\right\|_{\text{span}} + \frac{2}{n_i} + \frac{4\alpha(\delta,n_i)^2}{(1-\gamma_i^\star)^2 n_i} + \frac{8\alpha(\delta,n_i)}{1-\gamma_i^\star}\sqrt{\frac{2\left\|h^\star\right\|_{\text{span}} + 1 + \frac{1}{n_i}}{n_i}} + \frac{8}{n_i} \\
&\le 2\left\|h^\star\right\|_{\text{span}} + \frac{10}{n_i} + \frac{4\alpha(\delta,n_i)^2}{n_i} 4n_i\left(\left\|h^\star\right\|_{\text{span}} + 1\right) \\
&\quad + 8\alpha(\delta,n_i)2\sqrt{n_i\left(\left\|h^\star\right\|_{\text{span}} + 1\right)}\sqrt{\frac{2\left\|h^\star\right\|_{\text{span}} + 1 + \frac{1}{n_i}}{n_i}} \\
&\le 2\left\|h^\star\right\|_{\text{span}} + 5 + 16\alpha(\delta,n_i)^2\left(\left\|h^\star\right\|_{\text{span}} + 1\right) \\
&\quad + 16\alpha(\delta,n_i)\sqrt{n_i\left(\left\|h^\star\right\|_{\text{span}} + 1\right)}\sqrt{\frac{2\left\|h^\star\right\|_{\text{span}} + 2}{n_i}} \\
&= 2\left\|h^\star\right\|_{\text{span}} + 5 + 16\alpha(\delta,n_i)^2\left(\left\|h^\star\right\|_{\text{span}} + 1\right) + 16\sqrt{2}\alpha(\delta,n_i)\left(\left\|h^\star\right\|_{\text{span}} + 1\right) \\
&\le 44\alpha(\delta,n_i)^2\left(\left\|h^\star\right\|_{\text{span}} + 1\right)
\end{aligned} \tag{51}$$

where in the third inequality we used that $n_i \ge 2$ to bound $\frac{1}{n_i} \le \frac{1}{2} \le 1$, and in the final inequality we used that $\alpha(\delta, n_i) \ge 1$ (which is immediate from the form of α and the facts that $\delta \le 1, n_i \ge 1$) and that $5 + 16 + 16\sqrt{2} \le 44$. Substituting the inequality (51) into (49) and simplifying in the same

ways (including setting $\gamma = \gamma_i^\star$), we obtain that

$$\begin{aligned}
&\widehat{U}_i(\gamma_i^\star) - \widehat{L}_i(\gamma_i^\star) \\
&\leq (1-\gamma_i^\star)\left\|\widetilde{V}_{\gamma_i^\star,i}\right\|_{\text{span}} + 7\frac{1-\gamma_i^\star}{n_i} + \frac{2\alpha(\delta,n_i)^2}{(1-\gamma_i^\star)n_i} + 5\alpha(\delta,n_i)\sqrt{\frac{\left\|\widetilde{V}_{\gamma_i^\star,i}\right\|_{\text{span}} + 1 + \frac{3}{n_i}}{n_i}} \\
&\leq (1-\gamma_i^\star)44\alpha(\delta,n_i)^2\left(\|h^\star\|_{\text{span}}+1\right) + 7\frac{1-\gamma_i^\star}{n_i} + \frac{2\alpha(\delta,n_i)^2}{(1-\gamma_i^\star)n_i} \\
&\qquad + 5\alpha(\delta,n_i)\sqrt{\frac{44\alpha(\delta,n_i)^2\left(\|h^\star\|_{\text{span}}+1\right) + 1 + \frac{3}{n_i}}{n_i}} \\
&\leq \frac{1}{\sqrt{n_i\left(\|h^\star\|_{\text{span}}+1\right)}}44\alpha(\delta,n_i)^2\left(\|h^\star\|_{\text{span}}+1\right) + 7\frac{1}{n_i} + \frac{2\alpha(\delta,n_i)^2}{n_i}2\sqrt{n_i\left(\|h^\star\|_{\text{span}}+1\right)} \\
&\qquad + 5\alpha(\delta,n_i)\sqrt{\frac{47\alpha(\delta,n_i)^2\left(\|h^\star\|_{\text{span}}+1\right)}{n_i}} \\
&= \left(48+5\sqrt{47}\right)\alpha(\delta,n_i)^2\sqrt{\frac{\|h^\star\|_{\text{span}}+1}{n_i}} + \frac{7}{n_i} \\
&\leq 83\alpha(\delta,n_i)^2\sqrt{\frac{\|h^\star\|_{\text{span}}+1}{n_i}} + 7\sqrt{\frac{\|h^\star\|_{\text{span}}+1}{n_i}} \\
&\leq 90\alpha(\delta,n_i)^2\sqrt{\frac{\|h^\star\|_{\text{span}}+1}{n_i}}.
\end{aligned}$$

We have thus shown the first part of the lemma statement.

For the second part of the lemma statement, we need to find some i such that $90\alpha(\delta,n_i)^2\sqrt{\frac{\|h^\star\|_{\text{span}}+1}{n_i}} \leq \varepsilon$. First we compute

$$\begin{aligned}
\alpha(\delta,n_i)^4 &= 24^4 16^2 \log^2\left(\frac{24SAn_i^5}{\delta}\right)\log_2^4(\log_2(n_i+4)) \\
&\leq 24^4 16^2 117 \log^2\left(\frac{24SAn_i^5}{\delta}\right)\log(n_i) \\
&\leq 24^4 16^2 5^2 117 \log^2\left(\frac{24SAn_i}{\delta}\right)\log(n_i) \\
&\leq 24^4 16^2 5^2 117 \log^3\left(\frac{24SAn_i}{\delta}\right)
\end{aligned}$$

so

$$\begin{aligned}
&90\alpha(\delta,n_i)^2\sqrt{\frac{\|h^\star\|_{\text{span}}+1}{n_i}} \le \varepsilon \\
\iff & n_i \ge 90^2\alpha(\delta,n_i)^4\frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2} \\
\overset{\text{(I)}}{\impliedby} & n_i \ge 90^2 24^4 16^2 5^2 117\frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2}\log^3\left(\frac{24SAn_i}{\delta}\right) \\
\overset{\text{(II)}}{\impliedby} & n_i \ge 10\cdot 90^2 24^4 16^2 5^2 117\frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2}\log^3\left(\frac{240SA}{\delta}90^2 24^4 16^2 5^2 117\frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2}\right)
\end{aligned} \tag{52}$$

where (I) is due to Lemma 12 and (II) is due to Lemma 13. Therefore if we set

$$B = \left\lceil \log_2\left(C_1\frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2}\log^3\left(\frac{C_2SA(\|h^\star\|_{\text{span}}+1)}{\delta\varepsilon}\right)\right)\right\rceil$$

where $C_1 = 10\cdot 90^2 24^4 16^2 5^2 117$ and $C_2 = 240\cdot 90^2 24^4 16^2 5^2 117$, then (52) will be satisfied for $i = B$. ■

Lemma 12 $(\log_2\log_2(x+4))^4 \le 117\log x$ *for all* $x \ge 2$.

Proof Since $z \mapsto 3\cdot 2^{z/4}$ is convex, $3\cdot 2^{z/4}$ is lower-bounded by its tangent line at $z=4$, so for all z we have

$$3\cdot 2^{z/4} \ge 3\cdot 2^{4/4} + 3\cdot 2^{4/4}\cdot\ln(2^{1/4})\cdot(z-4) = 6 + \frac{3\ln(2)}{2}(z-4).$$

Since $\frac{3\ln(2)}{2} \in (1, 1.1)$, we have for all $z \ge 4$ that

$$6 + \frac{3\ln(2)}{2}(z-4) \ge 6 + (z-4) = z+2$$

and for all $0 \le z \le 4$ that

$$6 + \frac{3\ln(2)}{2}(z-4) \ge 6 + 1.1(z-4) > 1.1z + 1.5 \ge z + 1.5$$

so we have that $3\cdot 2^{z/4} \ge z + 1.5$ for all $z \ge 0$. Also $\log_2(1+\log_2 3) < 1.5$, so we have $3\cdot 2^{z/4} \ge z + \log_2(1+\log_2 3)$. Now letting $y = 2^z$ or equivalently $z = \log_2 y$ (and we must have $y \ge 1$ since $z \ge 0$), we have that

$$\begin{aligned}
\log_2(y + \log_2 3) \le \log_2(y + y\log_2 3) &= \log_2 y + \log_2(1+\log_2 3) \\
&= z + \log_2(1+\log_2 3) \le 3\cdot 2^{z/4} = 3y^{1/4}.
\end{aligned}$$

This implies

$$(\log_2(y+\log_2 3))^4 \le 3^4 y.$$

Now letting $x = 2^y$ or equivalently $y = \log_2(x)$ (and we must have $x \geq 2$ since $y \geq 1$), we have that

$$(\log_2(\log_2(x+4)))^4 \leq (\log_2(\log_2(3x)))^4 = (\log_2(\log_2 x + \log_2 3))^4 = (\log_2(y + \log_2 3))^4 \leq 3^4 y,$$

where the first inequality is because $x + 4 \leq 3x$ for $x \geq 2$. We also have

$$3^4 y = 3^4 \log_2(x) = 3^4 \log_2(e) \ln x < 117 \ln(x).$$

Thus, we can conclude the desired result for any $x \geq 2$ by making the appropriate choice of z (that is, $z = \log_2 \log_2 x$). ■

Lemma 13 *Suppose that $x, y \geq 1$. Then if $n \geq 10x \log^3(10xy)$, we have that $n \geq x \log^3(yn)$.*

Proof The desired conclusion $n \geq x \log^3(yn)$ is equivalent to $\frac{n}{\log^3(yn)} \geq x$. The derivative of $\frac{n}{\log^3(yn)}$ with respect to n is $\frac{\log(yn)-3}{\log^4(yn)}$ which is ≥ 0 if $n \geq \frac{e^3}{y}$, so the function $\frac{n}{\log^3(yn)}$ is monotone non-decreasing for $n \geq \frac{e^3}{y}$. Thus it suffices to find some m such that $m \geq \frac{e^3}{y}$ and $\frac{m}{\log^3(ym)} \geq x$, because then by monotonicity we have that $n \geq m$ implies $\frac{n}{\log^3(yn)} \geq \frac{m}{\log^3(ym)} \geq x$.

Now we claim that $m = 10x \log^3(10xy)$ satisfies these conditions. First, since $x, y \geq 1$, we have that $\log(10xy) > \log(e^2) = 2$, so $m = 10x \log^3(10xy) > 10 \cdot 2^3 = 80 \geq e^3 \geq \frac{e^3}{y}$, meeting the first required condition. Next, we have that

$$\begin{aligned}\log^3(ym) &= \log^3\left(y10x \log^3(10xy)\right) = (\log(10xy) + 3\log\log(10xy))^3 \\ &\leq \left(\log(10xy) + \frac{3}{e}\log(10xy)\right)^3 \leq 10\log^3(10xy)\end{aligned}$$

where we used that $\log x \leq \frac{x}{e}$ and then in the last step we used that $(1 + \frac{3}{e})^3 \leq 10$. (To obtain the inequality $\log x \leq \frac{x}{e}$, first we can show the inequality $x \leq e^{x/e}$ by noting that $e^{x/e}$ is convex, using the first-order convexity condition, and taking the tangent line to $e^{x/e}$ at $x = e$. Then we can take $\log$ of both sides.) Thus

$$\frac{m}{\log^3(ym)} \geq \frac{10x \log^3(10xy)}{10\log^3(10xy)} \geq x$$

as desired. ■

Combining the consequences of Lemmas 9 and 11, we can prove Theorem 2.

Proof of Theorem 2 For this proof we assume the event in Lemma 6 holds, which occurs with probability at least $1 - \delta$. First we argue that the algorithm terminates and uses at most the claimed number of samples. By Lemma 11, we have that $\widehat{U}_B(\gamma_B^\star) - \widehat{L}_B(\gamma_B^\star) \leq \varepsilon$. Since this would trigger the termination condition of Algorithm 2, this implies that the algorithm must terminate on or before

iteration B. Therefore the total number of samples used per state-action pair is at most

$$\begin{aligned}\sum_{i=1}^{B} n_i &= \sum_{i=1}^{B} 2^i \le 2^{B+1} \\ &= 2 \cdot 2^{\left\lceil \log_2\left(C_1 \frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2} \log^3\left(\frac{C_2 SA(\|h^\star\|_{\text{span}}+1)}{\delta\varepsilon}\right)\right)\right\rceil} \\ &\le 4C_1 \frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2} \log^3\left(\frac{C_2 SA(\|h^\star\|_{\text{span}}+1)}{\delta\varepsilon}\right) =: N.\end{aligned}$$

Also we can further bound B, which is an upper bound on the number of iterations, by

$$\begin{aligned}B &= \left\lceil \log_2\left(C_1 \frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2} \log^3\left(\frac{C_2 SA(\|h^\star\|_{\text{span}}+1)}{\delta\varepsilon}\right)\right)\right\rceil \\ &\le \log_2\left(C_1 \frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2} \log^3\left(\frac{C_2 SA(\|h^\star\|_{\text{span}}+1)}{\delta\varepsilon}\right)\right) + 1 \\ &= \log_2\left(2C_1 \frac{\|h^\star\|_{\text{span}}+1}{\varepsilon^2} \log^3\left(\frac{C_2 SA(\|h^\star\|_{\text{span}}+1)}{\delta\varepsilon}\right)\right) \\ &\le \log_2(N).\end{aligned}$$

Since the algorithm terminates and by definition of $\widehat{L}$ and $\widehat{U}$, it is immediate that we have $\widehat{U} - \widehat{L} \le \varepsilon$. By Lemma 9, we have for all $i \ge 1$ and all $\gamma \in \mathcal{H}_i$ that

$$\widehat{L}_i(\gamma)\mathbf{1} \le \rho^{\widetilde{\pi}_{\gamma,i}} \le \rho^\star \le \widehat{U}_i(\gamma)\mathbf{1}$$

so in particular we have, for the (random) final iteration I, that

$$\widehat{L}\mathbf{1} = \widehat{L}_I(\widehat{\gamma}_I)\mathbf{1} \le \rho^{\widetilde{\pi}_{\gamma,I}} = \rho^{\widehat{\pi}} \le \rho^\star \le \widehat{U}_I(\widehat{\gamma}_I)\mathbf{1} = \widehat{U}\mathbf{1}.$$

Finally, combining this with the fact that $\widehat{U} - \widehat{L} \le \varepsilon$ we see that

$$\rho^{\widehat{\pi}} \ge \widehat{L}\mathbf{1} \ge \widehat{U}\mathbf{1} - \varepsilon\mathbf{1} \ge \rho^\star - \varepsilon\mathbf{1}$$

as desired. ■

C.2. Proof of Theorem 1

We start by recalling the definitions of some objects which appear in Algorithm 1 for convenience. These definitions will be in effect for the entirety of this subsection. We define the empirical transition kernel $\widehat{P}(s' \mid s,a) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{S_{s,a}^i = s'\}$, for all $s' \in \mathcal{S}$, using the n samples drawn from all state-action pairs within Algorithm 1. Let $\mathcal{H} = \{\gamma : \text{there exists an integer } k \text{ such that } \sqrt{n} \le \frac{1}{1-\gamma} = 2^k \le n\}$. Then for all $\gamma \in \mathcal{H}$ we define the policy $\widetilde{\pi}_\gamma$ and value function $\widetilde{V}_\gamma$ as the outputs of SOLVEDMDP$(\widehat{P}, r, \gamma, \frac{1}{n})$.

Now let $\gamma^\star$ be the smallest member of $\mathcal{H}$ that corresponds to an effective horizon at least $\sqrt{n(\|h^\star\|_{\text{span}}+1)}$, that is,

$$\gamma^\star = \inf\left\{\gamma \in \mathcal{H} : \frac{1}{1-\gamma} \ge \sqrt{n\left(\|h^\star\|_{\text{span}}+1\right)}\right\}. \tag{53}$$

Lemma 14 *If $n \geq 4$ then $\mathcal{H}$ is nonempty. Furthermore if also $n \geq 16(\|h^\star\|_{\text{span}} + 1)$, then*

$$\sqrt{n(\|h^\star\|_{\text{span}} + 1)} \leq \frac{1}{1-\gamma^\star} \leq 2\sqrt{n(\|h^\star\|_{\text{span}} + 1)}. \tag{54}$$

Proof First we note that $\mathcal{H}$ is nonempty if the interval $[\sqrt{n}, n]$ contains some power of 2, which is ensured if $n \geq 2\sqrt{n}$, or equivalently if $n \geq 4$. It is also simple to check for $n \in \{1, 2, 3\}$ that $\mathcal{H}$ is also nonempty.

Next, the smallest power of 2 which is $\geq \sqrt{n(\|h^\star\|_{\text{span}} + 1)}$ is at most $2\sqrt{n(\|h^\star\|_{\text{span}} + 1)}$, so to guarantee it is contained in $\mathcal{H}$ we need it to be $\leq$ the largest element of $\mathcal{H}$, which is at least $n/2$. Therefore if

$$2\sqrt{n(\|h^\star\|_{\text{span}} + 1)} \leq n/2 \iff n \geq 16(\|h^\star\|_{\text{span}} + 1)$$

then (54) will be satisfied. ■

Lemma 15 *Define the function $\alpha(\tilde{\delta}, \tilde{n}) = 24\sqrt{16 \log\left(\frac{24SA\tilde{n}^5}{\tilde{\delta}}\right)} \log_2(\log_2(\tilde{n} + 4))$. Fix $\delta > 0$. Then with probability at least $1 - \delta$, we have for all $\gamma \in \mathcal{H}$ that*

$$\left\|V_\gamma^{\pi_\gamma^\star} - \widehat{V}_\gamma^{\pi_\gamma^\star}\right\|_\infty \leq \frac{\alpha(\delta, n)}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^{\pi_\gamma^\star}\right\|_{\text{span}} + 1}{n}} \tag{55}$$

and also the subroutine on line 7 outputs a policy $\widetilde{\pi}_\gamma$ such that

$$\widehat{V}_\gamma^{\widetilde{\pi}_\gamma} \geq \widehat{V}_\gamma^\star - \frac{1}{n}\mathbf{1} \tag{56}$$

$$\left\|\widetilde{V}_\gamma - \widehat{V}_\gamma^\star\right\|_\infty \leq \frac{1}{n} \tag{57}$$

$$\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma} - V_\gamma^{\widetilde{\pi}_\gamma}\right\|_\infty \leq \frac{\alpha(\delta, n)}{1-\gamma}\sqrt{\frac{\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}} + 1}{n}}. \tag{58}$$

Proof Following identical steps as to the proof of Lemma 6, but with n in place of n_i, up until equations (23) and (24), we obtain that with probability at least $1 - \delta$ the two inequalities

$$\left\|V_\gamma^{\pi_\gamma^\star} - \widehat{V}_\gamma^{\pi_\gamma^\star}\right\|_\infty \leq \frac{24\log_2\log_2(n+4)}{1-\gamma}\sqrt{16\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 1}{n}\log\left(\frac{24SAn^4}{\delta}\right)}$$

and

$$\left\|V_\gamma^{\widetilde{\pi}_\gamma} - \widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_\infty \leq \frac{24\log_2\log_2(n+4)}{1-\gamma}\sqrt{16\frac{\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}} + 1}{n}\log\left(\frac{24SAn^4}{\delta}\right)}$$

both hold. We obtain the desired conclusion by noting that

$$\alpha(\delta, n) \geq 24\log_2\log_2(n+4)\sqrt{16\log\left(\frac{24SAn^4}{\delta}\right)}.$$

■

Lemma 16 *Under the event described in Lemma 15, we have*

$$\rho^{\widetilde{\pi}_\gamma} \geq \widehat{L}(\gamma)\mathbf{1} = (1-\gamma)\min_s \widetilde{V}_\gamma(s)\mathbf{1} - 2\frac{1-\gamma}{n}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{\left\|\widetilde{V}_\gamma\right\|_{\text{span}} + \frac{3}{n} + 1}{n}}\mathbf{1} \tag{59}$$

for all $\gamma \in \mathcal{H}$.

Proof The proof is very similar to the first part of the proof of Lemma 9. Fix $\gamma \in \mathcal{H}$. Using inequality (15) from Lemma 5, then using the triangle inequality, then (56), then (58), then the triangle inequality again, then (57), we have

$$\begin{aligned}
\rho^{\widetilde{\pi}_\gamma} &\geq (1-\gamma)\min_s V_\gamma^{\widetilde{\pi}_\gamma}(s)\mathbf{1} \\
&\geq (1-\gamma)\min_s \widehat{V}_\gamma^{\widetilde{\pi}_\gamma}(s)\mathbf{1} - (1-\gamma)\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma} - V_\gamma^{\widetilde{\pi}_\gamma}\right\|_\infty \mathbf{1} \\
&\geq (1-\gamma)\min_s \widehat{V}_\gamma^{\star}(s)\mathbf{1} - \frac{1-\gamma}{n}\mathbf{1} - (1-\gamma)\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma} - V_\gamma^{\widetilde{\pi}_\gamma}\right\|_\infty \mathbf{1} \\
&\geq (1-\gamma)\min_s \widehat{V}_\gamma^{\star}(s)\mathbf{1} - \frac{1-\gamma}{n}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}} + 1}{n}}\mathbf{1} \\
&\geq (1-\gamma)\min_s \widetilde{V}_\gamma(s)\mathbf{1} - (1-\gamma)\left\|\widehat{V}_\gamma^{\star} - \widetilde{V}_\gamma\right\|_\infty \mathbf{1} - \frac{1-\gamma}{n}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}} + 1}{n}}\mathbf{1} \\
&\geq (1-\gamma)\min_s \widetilde{V}_\gamma(s)\mathbf{1} - 2\frac{1-\gamma}{n}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}} + 1}{n}}\mathbf{1}.
\end{aligned} \tag{60}$$

Now, nearly identically to the bound (43), we can use the requirements (56) and (57) to bound

$$\begin{aligned}
\left\|\widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}} &\leq \left\|\widehat{V}_\gamma^{\star}\right\|_{\text{span}} + \left\|\widehat{V}_\gamma^{\star} - \widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}} \\
&\leq \left\|\widetilde{V}_\gamma\right\|_{\text{span}} + \left\|\widetilde{V}_\gamma - \widehat{V}_\gamma^{\star}\right\|_{\text{span}} + \left\|\widehat{V}_\gamma^{\star} - \widehat{V}_\gamma^{\widetilde{\pi}_\gamma}\right\|_{\text{span}} \\
&\leq \left\|\widetilde{V}_\gamma\right\|_{\text{span}} + 2\left\|\widetilde{V}_\gamma - \widehat{V}_\gamma^{\star}\right\|_\infty + \frac{1}{n} \\
&\leq \left\|\widetilde{V}_\gamma\right\|_{\text{span}} + \frac{3}{n}.
\end{aligned} \tag{61}$$

Finally combining (61) with (60), we obtain that

$$\rho^{\widetilde{\pi}_\gamma} \geq (1-\gamma)\min_s \widetilde{V}_\gamma(s)\mathbf{1} - 2\frac{1-\gamma}{n}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{\left\|\widetilde{V}_\gamma\right\|_{\text{span}} + \frac{3}{n} + 1}{n}}\mathbf{1} = \widehat{L}(\gamma)\mathbf{1}$$

as desired. ■

Lemma 17 *Under the event in Lemma 15, for all $\gamma \in \mathcal{H}$, we have*

$$\left\|\widehat{V}_\gamma^\star - V_\gamma^\star\right\|_\infty \le \frac{2\alpha(\delta,n)^2}{(1-\gamma)^2 n} + \frac{4\alpha(\delta,n)}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\star\right\|_{\text{span}} + 1 + \frac{1}{n}}{n}} + \frac{4}{n}. \tag{62}$$

Proof Similarly to Lemma 8, this follows immediately from Lemma 7 as its conditions are satisfied under the event in Lemma 15 (by setting $m = n$ and $\beta = \alpha(\delta, n)$ for all $\gamma \in \mathcal{H}$). ■

Lemma 18 *Suppose $n \ge 16(\|h^\star\|_{\text{span}} + 1)$. Under the event described in Lemma 15, we have*

$$\widehat{L}(\gamma^\star)\mathbf{1} \ge \rho^\star - 30\alpha(\delta,n)^2\sqrt{\frac{\|h^\star\|_{\text{span}} + 1}{n}}\mathbf{1}.$$

Proof First we bound $\left\|\widehat{V}_\gamma^\star - V_\gamma^\star\right\|_\infty$ in terms of $\|h^\star\|_{\text{span}}$. Using the bound (62) from Lemma 17 and substituting $\gamma = \gamma^\star$, as well as using Lemma 14 to bound $\frac{1}{1-\gamma^\star}$, we have

$$\begin{aligned}
\left\|\widehat{V}_{\gamma^\star}^\star - V_{\gamma^\star}^\star\right\|_\infty &\le \frac{2\alpha(\delta,n)^2}{(1-\gamma^\star)^2 n} + \frac{4\alpha(\delta,n)}{1-\gamma^\star}\sqrt{\frac{\left\|V_{\gamma^\star}^\star\right\|_{\text{span}} + 1 + \frac{1}{n}}{n}} + \frac{4}{n} \\
&\le \frac{2\alpha(\delta,n)^2}{n} 4n\left(\|h^\star\|_{\text{span}} + 1\right) + 8\alpha(\delta,n)\sqrt{n\left(\|h^\star\|_{\text{span}} + 1\right)}\sqrt{\frac{\left\|V_{\gamma^\star}^\star\right\|_{\text{span}} + 1 + \frac{1}{n}}{n}} + \frac{4}{n} \\
&\le 8\alpha(\delta,n)^2\left(\|h^\star\|_{\text{span}} + 1\right) + 8\alpha(\delta,n)\sqrt{n\left(\|h^\star\|_{\text{span}} + 1\right)}\sqrt{\frac{2\|h^\star\|_{\text{span}} + 2}{n}} + \frac{1}{4} \\
&\le 8\alpha(\delta,n)^2\left(\|h^\star\|_{\text{span}} + 1\right) + (8\sqrt{2} + 1/4)\alpha(\delta,n)\left(\|h^\star\|_{\text{span}} + 1\right) \\
&\le 20\alpha(\delta,n)^2\left(\|h^\star\|_{\text{span}} + 1\right)
\end{aligned} \tag{63}$$

also using the facts that $n \ge 16$, that $\alpha(\delta, n) \ge 1$, and that $\left\|V_{\gamma^\star}^\star\right\|_{\text{span}} \le 2\|h^\star\|_{\text{span}}$ (Wei et al., 2020, Lemma 2).

We can use the triangle inequality, that $\|\cdot\|_{\text{span}} \le 2\|\cdot\|_\infty$, triangle inequality again, that $\left\|V_{\gamma^\star}^\star\right\|_{\text{span}} \le 2\|h^\star\|_{\text{span}}$, (57) and (63), and then that $\alpha(\delta, n) \ge 1$ and $n \ge 16$ to bound

$$\begin{aligned}
\left\|\widetilde{V}_{\gamma^\star}\right\|_{\text{span}} &\le \left\|V_{\gamma^\star}^\star\right\|_{\text{span}} + \left\|\widetilde{V}_{\gamma^\star} - V_{\gamma^\star}^\star\right\|_{\text{span}} \\
&\le \left\|V_{\gamma^\star}^\star\right\|_{\text{span}} + 2\left\|\widetilde{V}_{\gamma^\star} - V_{\gamma^\star}^\star\right\|_\infty \\
&\le 2\|h^\star\|_{\text{span}} + 2\left\|\widetilde{V}_{\gamma^\star} - \widehat{V}_{\gamma^\star}^\star\right\|_\infty + 2\left\|\widehat{V}_{\gamma^\star}^\star - V_{\gamma^\star}^\star\right\|_\infty \\
&\le 2\|h^\star\|_{\text{span}} + \frac{2}{n} + 40\alpha(\delta,n)^2\left(\|h^\star\|_{\text{span}} + 1\right) \\
&\le 42\alpha(\delta,n)^2\left(\|h^\star\|_{\text{span}} + 1\right).
\end{aligned} \tag{64}$$

Now using the inequalities (63) and (64) to lower-bound $\widehat{L}(\gamma^\star)$, we obtain

$$
\begin{aligned}
\widehat{L}(\gamma^\star)\mathbf{1} &= (1-\gamma^\star)\min_s \widetilde{V}_{\gamma^\star}(s)\mathbf{1} - 2\frac{1-\gamma^\star}{n}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{\left\|\widetilde{V}_{\gamma^\star}\right\|_{\text{span}} + \frac{3}{n} + 1}{n}}\mathbf{1} \\
&\geq (1-\gamma^\star)\min_s \widehat{V}^\star_{\gamma^\star}(s)\mathbf{1} - 3\frac{1-\gamma^\star}{n}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{\left\|\widetilde{V}_{\gamma^\star}\right\|_{\text{span}} + \frac{3}{n} + 1}{n}}\mathbf{1} \\
&\geq (1-\gamma^\star)\min_s V^\star_{\gamma^\star}(s)\mathbf{1} - (1-\gamma^\star)\left\|\widehat{V}^\star_{\gamma^\star} - V^\star_{\gamma^\star}\right\|_\infty \mathbf{1} - 3\frac{1-\gamma^\star}{n}\mathbf{1} \\
&\quad - \alpha(\delta,n)\sqrt{\frac{\left\|\widetilde{V}_{\gamma^\star}\right\|_{\text{span}} + \frac{3}{n} + 1}{n}}\mathbf{1} \\
&\geq (1-\gamma^\star)\min_s V^\star_{\gamma^\star}(s)\mathbf{1} - (1-\gamma^\star)21\alpha(\delta,n)^2\left(\|h^\star\|_{\text{span}} + 1\right)\mathbf{1} - 3\frac{1-\gamma^\star}{n}\mathbf{1} \\
&\quad - \alpha(\delta,n)^2\sqrt{\frac{42\left(\|h^\star\|_{\text{span}} + 1\right) + \frac{3}{n} + 1}{n}}\mathbf{1} \\
&\geq \rho^\star - 2(1-\gamma^\star)\|h^\star\|_{\text{span}}\mathbf{1} - (1-\gamma^\star)21\alpha(\delta,n)^2\left(\|h^\star\|_{\text{span}} + 1\right)\mathbf{1} - 3\frac{1-\gamma^\star}{n}\mathbf{1} \\
&\quad - \alpha(\delta,n)^2\sqrt{\frac{44\left(\|h^\star\|_{\text{span}} + 1\right)}{n}}\mathbf{1} \\
&\geq \rho^\star - (1-\gamma^\star)22\alpha(\delta,n)^2\left(\|h^\star\|_{\text{span}} + 1\right)\mathbf{1} - \alpha(\delta,n)^2\sqrt{\frac{44\left(\|h^\star\|_{\text{span}} + 1\right)}{n}}\mathbf{1} \\
&\geq \rho^\star - \frac{1}{\sqrt{n\left(\|h^\star\|_{\text{span}} + 1\right)}}22\alpha(\delta,n)^2\left(\|h^\star\|_{\text{span}} + 1\right)\mathbf{1} - \alpha(\delta,n)^2\sqrt{\frac{44\left(\|h^\star\|_{\text{span}} + 1\right)}{n}}\mathbf{1} \\
&\geq \rho^\star - 30\alpha(\delta,n)^2\sqrt{\frac{\|h^\star\|_{\text{span}} + 1}{n}}\mathbf{1}
\end{aligned}
$$

where in the first inequality we used that $\left\|\widehat{V}^\star_{\gamma^\star} - \widetilde{V}_{\gamma^\star}\right\|_\infty \leq \frac{1}{n}$ from (57) and triangle inequality, in the second inequality we used triangle inequality, in the third we used (63) and (64), in the fourth we used that $\min_s V^\star_{\gamma^\star}(s) \geq \frac{1}{1-\gamma^\star}\rho^\star - \|h^\star\|_{\text{span}}$ from (Wei et al., 2020, Lemma 2) and that $n \geq 16$, in the fifth we again used that $n \geq 16$, and in the sixth we used (54) to upper bound $(1-\gamma^\star)$. ■

Now we combine these intermediate results to prove Theorem 1.

Proof of Theorem 1 Under the event in Lemma 15 which holds with probability at least $1-\delta$, we have by Lemma 16 that $\rho^{\widetilde{\pi}_{\widehat{\gamma}}} \geq \widehat{L}(\widehat{\gamma})\mathbf{1}$. Since we trivially have $\rho^{\widetilde{\pi}_{\widehat{\gamma}}} \geq 0\mathbf{1}$, this implies $\rho^{\widetilde{\pi}_{\widehat{\gamma}}} \geq \max\{\widehat{L}(\widehat{\gamma}), 0\}\mathbf{1} = \widehat{\rho}$. Now to lower bound $\widehat{\rho}$ we consider two cases. First, if $n \geq 16(\|h^\star\|_{\text{span}} + 1)$,

then we have

$$\widehat{\rho} = \max\{\widehat{L}(\widehat{\gamma}), 0\}\mathbf{1} \geq \widehat{L}(\widehat{\gamma}) \geq \widehat{L}(\gamma^\star)\mathbf{1} \geq \rho^\star - 30\alpha(\delta, n)^2\sqrt{\frac{\|h^\star\|_{\text{span}} + 1}{n}}\mathbf{1}$$

where the second inequality step is by the definition of $\widehat{\gamma}$, and third inequality is from Lemma 18. Next, if $n < 16(\|h^\star\|_{\text{span}} + 1)$, then

$$\widehat{\rho} = \max\{\widehat{L}(\widehat{\gamma}), 0\}\mathbf{1} \geq 0\mathbf{1} \geq \rho^\star - 30\alpha(\delta, n)^2\sqrt{\frac{\|h^\star\|_{\text{span}} + 1}{n}}\mathbf{1}$$

where the final inequality is because $\rho^\star \leq \mathbf{1}$ so the condition on n ensures the RHS is $< \mathbf{0}$ (note $\alpha(\delta, n) \geq 1$). Finally, we let $C_3 = 30$. ■

C.3. Proof of Lemma 3

Proof of Lemma 3 First we check that T ensures that $\gamma^T \leq \frac{(1-\gamma)^2\varepsilon}{3}$. We have

$$T \geq \frac{\log(\frac{3}{(1-\gamma)^2\varepsilon})}{1-\gamma} \geq \frac{\log(\frac{3}{(1-\gamma)^2\varepsilon})}{\log(1/\gamma)}$$

using the fact that $\frac{1}{1-\gamma} \geq \frac{1}{\log(1/\gamma)}$ for $\gamma \in (0, 1)$. Thus

$$\gamma^T \leq \gamma^{\frac{\log(\frac{3}{(1-\gamma)^2\varepsilon})}{\log(1/\gamma)}} = \left(e^{-1}\right)^{\log(\frac{3}{(1-\gamma)^2\varepsilon})} = \frac{(1-\gamma)^2\varepsilon}{3}. \tag{65}$$

By (Fruit et al., 2018, Lemma 16) Clip_M is non-expansive with respect to $\|\cdot\|_\infty$, so since $\mathcal{T}_\gamma$ is γ-contractive, we have that $\mathcal{L} = \text{Clip}_M \circ \mathcal{T}_\gamma$ is also γ-contractive. It is a standard fact that this implies $\mathcal{L}$ has a unique fixed point, which we name $V^\star_{\gamma,M}$.

Thus

$$\begin{aligned}\left\|V^T - V^\star_{\gamma,M}\right\|_\infty &= \left\|\mathcal{L}(V^{T-1}) - \mathcal{L}(V^\star_{\gamma,M})\right\|_\infty \\ &\leq \gamma\left\|V^{T-1} - V^\star_{\gamma,M}\right\|_\infty \leq \cdots \leq \gamma^T\left\|V^0 - V^\star_{\gamma,M}\right\|_\infty \leq \gamma^T\frac{1}{1-\gamma},\end{aligned}$$

which is $\leq \frac{(1-\gamma)\varepsilon}{3} \leq \varepsilon$ using (65). Similarly, we have

$$\begin{aligned}\left\|V^T - \mathcal{L}(V^T)\right\|_\infty = \left\|\mathcal{L}(V^{T-1}) - \mathcal{L}(V^T)\right\|_\infty &\leq \gamma\left\|V^{T-1} - V^T\right\|_\infty = \gamma\left\|V^{T-1} - \mathcal{L}(V^{T-1})\right\|_\infty \\ &\leq \cdots \leq \gamma^T\left\|V^0 - \mathcal{L}(V^0)\right\|_\infty \leq \gamma^T\frac{1}{1-\gamma} \leq \frac{(1-\gamma)\varepsilon}{3}\end{aligned}$$

again using (65). By definition of $\tilde{r}$, we immediately have that $\tilde{r} \leq r$. Also by construction of $\tilde{r}$ we have that

$$\tilde{r}_{\widehat{\pi}} + \gamma P_{\widehat{\pi}} V^T = \min\left\{r_{\widehat{\pi}} + \gamma P_{\widehat{\pi}} V^T, M\mathbf{1} + \min_{s'} V^T(s')\mathbf{1}\right\},$$

where the min is elementwise. Thus

$$\begin{aligned}
&\left\|\mathcal{L}(V^T) - \tilde{r}_{\widehat{\pi}} - \gamma P_{\widehat{\pi}} V^T\right\|_\infty \\
&= \left\|\min\left\{r_{\widehat{\pi}} + \gamma P_{\widehat{\pi}} V^T, M\mathbf{1} + \min_{s'}(r_{\widehat{\pi}} + \gamma P_{\widehat{\pi}} V^T)(s')\mathbf{1}\right\} - \min\left\{r_{\widehat{\pi}} + \gamma P_{\widehat{\pi}} V^T, M\mathbf{1} + \min_{s'} V^T(s')\mathbf{1}\right\}\right\|_\infty \\
&\leq \left\|M\mathbf{1} + \min_{s'}(r_{\widehat{\pi}} + \gamma P_{\widehat{\pi}} V^T)(s')\mathbf{1} - \left(M\mathbf{1} + \min_{s'} V^T(s')\mathbf{1}\right)\right\|_\infty \\
&= \left|\min_{s'}(r_{\widehat{\pi}} + \gamma P_{\widehat{\pi}} V^T)(s') - \min_{s'} V^T(s')\right| \\
&= \left|\min_{s'} \mathrm{Clip}_M(r_{\widehat{\pi}} + \gamma P_{\widehat{\pi}} V^T)(s') - \min_{s'} V^T(s')\right| \\
&= \left|\min_{s'} \mathrm{Clip}_M(\mathcal{T}_\gamma(V^T))(s') - \min_{s'} V^T(s')\right| \\
&= \left|\min_{s'} \mathcal{L}(V^T)(s') - \min_{s'} V^T(s')\right| \\
&\leq \left\|\mathcal{L}(V^T) - V^T\right\|_\infty, \qquad (66)
\end{aligned}$$

where we use the elementary fact that $|\min\{a,b\} - \min\{c,d\}| \leq \max\{|a-c|, |b-d|\}$ for $a, b, c, d \in \mathbb{R}$, as well as the fact that Clip_M does not change the minimum entry of its input.

Now we can calculate that

$$\begin{aligned}
\left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - V^T\right\|_\infty &\leq \left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - \mathcal{L}(V^T)\right\|_\infty + \left\|\mathcal{L}(V^T) - V^T\right\|_\infty \\
&\leq \left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - \tilde{r}_{\widehat{\pi}} - \gamma P_{\widehat{\pi}} V^T\right\|_\infty + \left\|\mathcal{L}(V^T) - \tilde{r}_{\widehat{\pi}} - \gamma P_{\widehat{\pi}} V^T\right\|_\infty + \left\|\mathcal{L}(V^T) - V^T\right\|_\infty \\
&= \left\|\tilde{r}_{\widehat{\pi}} - \gamma P_{\widehat{\pi}} V^{\widehat{\pi}}_{\gamma,\tilde{r}} - \tilde{r}_{\widehat{\pi}} - \gamma P_{\widehat{\pi}} V^T\right\|_\infty + \left\|\mathcal{L}(V^T) - \tilde{r}_{\widehat{\pi}} - \gamma P_{\widehat{\pi}} V^T\right\|_\infty + \left\|\mathcal{L}(V^T) - V^T\right\|_\infty \\
&\overset{\text{(I)}}{\leq} \left\|\tilde{r}_{\widehat{\pi}} - \gamma P_{\widehat{\pi}} V^{\widehat{\pi}}_{\gamma,\tilde{r}} - \tilde{r}_{\widehat{\pi}} - \gamma P_{\widehat{\pi}} V^T\right\|_\infty + 2\left\|\mathcal{L}(V^T) - V^T\right\|_\infty \\
&\overset{\text{(II)}}{\leq} \gamma\left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - V^T\right\|_\infty + 2\left\|\mathcal{L}(V^T) - V^T\right\|_\infty,
\end{aligned}$$

where in (I) we used (66) and in (II) we used that $\|P_{\widehat{\pi}}\|_{\infty\to\infty} \leq 1$. Thus by rearranging we have that

$$\left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - V^T\right\|_\infty \leq \frac{2\left\|\mathcal{L}(V^T) - V^T\right\|_\infty}{1-\gamma} \leq \frac{2}{3}\varepsilon$$

using (65). Then we can bound

$$\left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - V^\star_{\gamma,M}\right\|_\infty \leq \left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - V^T\right\|_\infty + \left\|V^T - V^\star_{\gamma,M}\right\|_\infty \leq \frac{2}{3}\varepsilon + (1-\gamma)\frac{\varepsilon}{3} \leq \varepsilon$$

as desired. Also, $\left\|V^\star_{\gamma,M}\right\|_{\text{span}} \leq M$ since it is equal to $\mathrm{Clip}_M(\mathcal{T}_\gamma(V^\star_{\gamma,M}))$ and the output of Clip_M clearly has span bounded by M. Therefore

$$\left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}}\right\|_{\text{span}} \leq \left\|V^\star_{\gamma,M}\right\|_{\text{span}} + \left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - V^\star_{\gamma,M}\right\|_{\text{span}} \leq \left\|V^\star_{\gamma,M}\right\|_{\text{span}} + 2\left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - V^\star_{\gamma,M}\right\|_\infty \leq M + 2\varepsilon.$$

Finally, letting $\mathcal{T}^{\pi'}_{\gamma,r'}$ be the Bellman consistency/evaluation operator for the policy π' in the DMDP (P, r', γ), we have $\mathcal{T}^{\pi'}_{\gamma,r'}(V^{\pi'}_{\gamma,r'}) = V^{\pi'}_{\gamma,r'}$. Also, since $\left\|V^{\pi'}_{\gamma,r'}\right\|_{\text{span}} \leq M$, we have $\text{Clip}_M(\mathcal{T}^{\pi'}_{\gamma,r'}(V^{\pi'}_{\gamma,r'})) = V^{\pi'}_{\gamma,r'}$. Letting $\mathcal{L}^{\pi'} = \text{Clip}_M \circ \mathcal{T}^{\pi'}_{\gamma,r'}$, we have that $\mathcal{L}^{\pi'}$ is a γ contraction with respect to $\|\cdot\|_\infty$ (since as discussed above, Clip_M is $\|\cdot\|_\infty$-nonexpansive, and because $\mathcal{T}^{\pi'}_{\gamma,r'}$ is well-known to be γ-contractive with respect to $\|\cdot\|_\infty$). Thus $\mathcal{L}^{\pi'}$ has a unique fixed point, which must be $V^{\pi'}_{\gamma,r'}$ (since we have already verified $V^{\pi'}_{\gamma,r'}$ is a fixed point), and furthermore Picard iteration will converge to this fixed point. The operator Clip_M is also monotonic (in the sense that for any $x, y \in \mathbb{R}^{\mathcal{S}}$, $x \leq y \implies \text{Clip}_M(x) \leq \text{Clip}_M(y)$) as shown in (Fruit et al., 2018, Lemma 16). It is well-known that $\mathcal{T}_\gamma$ and $\mathcal{T}^{\pi'}_{\gamma,r'}$ are also monotonic, which immediately implies that the compositions $\mathcal{L}$ and $\mathcal{L}^{\pi'}$ are monotonic. It is immediate that for any $x \in \mathbb{R}^{\mathcal{S}}$ we have $\mathcal{T}^{\pi'}_{\gamma,r'}(x) \leq \mathcal{T}_\gamma(x)$, since for any $x \in \mathbb{R}^{\mathcal{S}}$ we have

$$\mathcal{T}^{\pi'}_{\gamma,r'}(x) = M^\pi(r' + \gamma P x) \leq M^\pi(r + \gamma P x) \leq M(r + \gamma P x) = \mathcal{T}_\gamma(x)$$

crucially using the fact that $r' \leq r$ and also monotonicity of M^π. This thus implies that $\mathcal{L}^{\pi'}(x) = \text{Clip}_M(\mathcal{T}^{\pi'}_{\gamma,r'}(x)) \leq \text{Clip}_M(\mathcal{T}_\gamma(x)) \leq \mathcal{L}(x)$ for any $x \in \mathbb{R}^{\mathcal{S}}$ (using monotonicity of Clip_M). Therefore $\mathcal{L}^{\pi'}(\mathbf{0}) \leq \mathcal{L}(\mathbf{0})$, and if it holds for some integer $i \geq 1$ that $(\mathcal{L}^{\pi'})^{(i)}(\mathbf{0}) \leq \mathcal{L}^{(i)}(\mathbf{0})$ then we can use the above-discussed properties to obtain that

$$(\mathcal{L}^{\pi'})^{(i+1)}(\mathbf{0}) = \mathcal{L}^{\pi'}\left((\mathcal{L}^{\pi'})^{(i)}(\mathbf{0})\right) \leq \mathcal{L}\left((\mathcal{L}^{\pi'})^{(i)}(\mathbf{0})\right) \leq \mathcal{L}\left(\mathcal{L}^{(i)}(\mathbf{0})\right) = \mathcal{L}^{(i+1)}(\mathbf{0}),$$

so we have by induction that $(\mathcal{L}^{\pi'})^{(i)}(\mathbf{0}) \leq \mathcal{L}^{(i)}(\mathbf{0})$ holds for all i, and thus

$$V^{\pi'}_{\gamma,r'} = \lim_{i\to\infty} (\mathcal{L}^{\pi'})^{(i)}(\mathbf{0}) \leq \lim_{i\to\infty} \mathcal{L}^{(i)}(\mathbf{0}) = V^\star_{\gamma,M}.$$

Thus we have

$$V^{\widehat{\pi}}_\gamma \geq V^{\widehat{\pi}}_{\gamma,\tilde{r}} \geq V^\star_{\gamma,M} - \varepsilon\mathbf{1} \geq V^{\pi'}_{\gamma,r'} - \varepsilon\mathbf{1}$$

as desired, where the first inequality is because $r \geq \tilde{r}$, the second inequality uses that $\left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - V^\star_{\gamma,M}\right\|_\infty \leq \varepsilon$, and the final inequality uses the above argument. ∎

C.4. Proof of Theorem 4

First we develop bounds on the error between value functions within P and $\widehat{P}$.

Lemma 19 *Fix a policy π, $\gamma \in (0,1)$, and $\delta, n > 0$. Then with probability at least $1-\delta$, we have that*

$$\left\|V^\pi_\gamma - \widehat{V}^\pi_\gamma\right\|_\infty \leq \frac{24\log_2\log_2(\frac{1}{1-\gamma}+4)}{1-\gamma}\sqrt{\frac{\left\|V^\pi_\gamma\right\|_{\text{span}}+1}{n}16\log\left(\frac{12SAn}{(1-\gamma)^2\delta}\right)}.$$

Proof We can reuse the proof of (Zurek and Chen, 2024, Theorem 9). Specifically, (Zurek and Chen, 2024, Equation 31) is stated for an optimal policy in the DMDP (P, r, γ), but completely identical

arguments actually hold for any fixed policy π. Therefore (Zurek and Chen, 2024, Equation 31) yields that with probability at least $1-\delta$ it holds that

$$\left\|V_\gamma^\pi - \widehat{V}_\gamma^\pi\right\|_\infty \le \frac{24\log_2\log_2(\frac{1}{1-\gamma}+4)}{1-\gamma}\sqrt{\frac{\left\|V_\gamma^\pi\right\|_{\text{span}}+1}{n}16\log\left(\frac{12SAn}{(1-\gamma)^2\delta}\right)}.$$

This completes the proof. ■

We would like to have a similar bound involving policies output by the span-constrained planning procedure applied to $\widehat{P}$, but since such policies are probabilistically dependent on $\widehat{P}$, much more effort is needed. In the absence of span-constrained planning, Agarwal et al. (2020) introduce the absorbing MDP construction to overcome such issues, and our approach is strongly inspired by theirs, although ours requires new modifications which are specific to the span-constrained planning procedure Algorithm 3.

First we must define the key quantities involved in our construction. Fix $s \in \mathcal{S}$ and $u \in [0,1]$. We define the MDP transition kernel $\widehat{P}^{(s)}$ as

$$\widehat{P}^{(s)}(s' \mid s'', a) = \begin{cases} \widehat{P}(s' \mid s'', a) & s'' \neq s \\ 1 & s'' = s' = s \\ 0 & s'' = s \text{ and } s' \neq s \end{cases}.$$

In words, $\widehat{P}^{(s)}$ is identical to $\widehat{P}$ except state s is made to be absorbing. We also define the reward function $r^{(s,u)}$ as

$$r^{(s,u)}(s', a) = \begin{cases} r(s', a) & s' \neq s \\ u & s' = s \end{cases}.$$

In words, $r^{(s,u)}$ is identical to r except state s gives reward u for all actions. For any $\gamma \in (0,1)$ and any $M > 0$, let $\widehat{\mathcal{T}}_\gamma^{(s,u)} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ be the γ-discounted Bellman operator for the DMDP $(\widehat{P}^{(s)}, r^{(s,u)}, \gamma)$, that is,

$$\widehat{\mathcal{T}}_\gamma^{(s,u)}(x) = M(r^{(s,u)} + \gamma \widehat{P}^{(s)} x)$$

for all $x \in \mathbb{R}^{\mathcal{S}}$. Let $\widehat{V}_{\gamma,M}^{(s,u)}$ be the unique fixed point of $\text{Clip}_M \circ \widehat{\mathcal{T}}_\gamma^{(s,u)}$ (this fixed point exists and is unique because $\text{Clip}_M \circ \widehat{\mathcal{T}}_\gamma^{(s,u)}$ is a γ-contraction due to Lemma 3).

Now we can summarize the key properties of this construction in the following lemma.

Lemma 20 *Fixing $\gamma \in (0,1)$ and $M > 0$, for any $u, u' \in [0,1]$ we have*

$$\left\|\widehat{V}_{\gamma,M}^{(s,u)} - \widehat{V}_{\gamma,M}^{(s,u')}\right\|_\infty \le \frac{|u-u'|}{1-\gamma}.$$

Also, letting $u^\star(s) = \widehat{\mathcal{T}}_\gamma(\widehat{V}_{\gamma,M}^\star)(s) - \gamma \widehat{V}_{\gamma,M}^\star(s)$, we have that $u^\star(s) \in [0,1]$ and

$$\widehat{V}_{\gamma,M}^{(s,u^\star(s))} = \widehat{V}_{\gamma,M}^\star,$$

where $\widehat{V}^{\star}_{\gamma,M}$ *is the unique fixed point of* $\mathrm{Clip}_M \circ \widehat{\mathcal{T}}_\gamma$.

Consequently, there exists a finite set U *with* $|U| = \left\lceil \frac{1}{2(1-\gamma)\varepsilon} \right\rceil$ *such that almost surely, for any* $s \in \mathcal{S}$ *there exists* $u \in U$ *such that* $\left\| \widehat{V}^{\star}_{\gamma,M} - \widehat{V}^{(s,u)}_{\gamma,M} \right\|_\infty \leq \varepsilon$.

Lastly, for any $u \in [0,1], s \in \mathcal{S}$, *and* $a \in \mathcal{A}$, $\widehat{V}^{(s,u)}_{\gamma,M}$ *is independent of the samples* $S^1_{s,a}, \ldots, S^n_{s,a}$ *used to construct* $\widehat{P}(\cdot \mid s,a)$.

Proof For the first inequality, we can calculate that

$$\begin{aligned}
\left\| \widehat{V}^{(s,u)}_{\gamma,M} - \widehat{V}^{(s,u')}_{\gamma,M} \right\|_\infty &= \left\| \mathrm{Clip}_M \left(\widehat{\mathcal{T}}^{(s,u)}_\gamma \left(\widehat{V}^{(s,u)}_{\gamma,M} \right) \right) - \mathrm{Clip}_M \left(\widehat{\mathcal{T}}^{(s,u')}_\gamma \left(\widehat{V}^{(s,u')}_{\gamma,M} \right) \right) \right\|_\infty \\
&\leq \left\| \widehat{\mathcal{T}}^{(s,u)}_\gamma \left(\widehat{V}^{(s,u)}_{\gamma,M} \right) - \widehat{\mathcal{T}}^{(s,u')}_\gamma \left(\widehat{V}^{(s,u')}_{\gamma,M} \right) \right\|_\infty \\
&= \left\| M \left(r^{(s,u)} + \gamma \widehat{P}^{(s)} \widehat{V}^{(s,u)}_{\gamma,M} \right) - M \left(r^{(s,u')} + \gamma \widehat{P}^{(s)} \widehat{V}^{(s,u')}_{\gamma,M} \right) \right\|_\infty \\
&\leq \left\| r^{(s,u)} - r^{(s,u')} + \gamma \widehat{P}^{(s)} \left(\widehat{V}^{(s,u)}_{\gamma,M} - \widehat{V}^{(s,u')}_{\gamma,M} \right) \right\|_\infty \\
&\leq \left\| r^{(s,u)} - r^{(s,u')} \right\|_\infty + \gamma \left\| \widehat{V}^{(s,u)}_{\gamma,M} - \widehat{V}^{(s,u')}_{\gamma,M} \right\|_\infty \\
&= |u - u'| + \gamma \left\| \widehat{V}^{(s,u)}_{\gamma,M} - \widehat{V}^{(s,u')}_{\gamma,M} \right\|_\infty .
\end{aligned}$$

Rearranging, we obtain that

$$\left\| \widehat{V}^{(s,u)}_{\gamma,M} - \widehat{V}^{(s,u')}_{\gamma,M} \right\|_\infty \leq \frac{|u - u'|}{1-\gamma}$$

as desired.

Now we show the second statement in the lemma. First we show that $\widehat{V}^{(s,u^\star(s))}_{\gamma,M} = \widehat{V}^{\star}_{\gamma,M}$. To show this, it suffices to show that

$$\widehat{\mathcal{T}}_\gamma \left(\widehat{V}^{\star}_{\gamma,M} \right) = \widehat{\mathcal{T}}^{(s,u^\star(s))}_\gamma \left(\widehat{V}^{\star}_{\gamma,M} \right), \tag{67}$$

since this implies that

$$\widehat{V}^{\star}_{\gamma,M} = \mathrm{Clip}_M \left(\widehat{\mathcal{T}}_\gamma \left(\widehat{V}^{\star}_{\gamma,M} \right) \right) = \mathrm{Clip}_M \left(\widehat{\mathcal{T}}^{(s,u^\star(s))}_\gamma \left(\widehat{V}^{\star}_{\gamma,M} \right) \right)$$

(using the fact that $\widehat{V}^{\star}_{\gamma,M}$ is a fixed point of $\mathrm{Clip}_M \circ \widehat{\mathcal{T}}_\gamma$ in the first equality), meaning that $\widehat{V}^{\star}_{\gamma,M}$ is a fixed point of $\mathrm{Clip}_M \circ \widehat{\mathcal{T}}^{(s,u^\star(s))}_\gamma$, and since the unique fixed point of $\mathrm{Clip}_M \circ \widehat{\mathcal{T}}^{(s,u^\star(s))}_\gamma$ is $\widehat{V}^{(s,u^\star(s))}_{\gamma,M}$, this would imply that $\widehat{V}^{(s,u^\star(s))}_{\gamma,M} = \widehat{V}^{\star}_{\gamma,M}$. Now focusing on (67), we note that by construction of $\widehat{\mathcal{T}}^{s,u^\star(s)}_\gamma$ we already have $\widehat{\mathcal{T}}_\gamma \left(\widehat{V}^{\star}_{\gamma,M} \right)(s') = \widehat{\mathcal{T}}^{(s,u^\star(s))}_\gamma \left(\widehat{V}^{\star}_{\gamma,M} \right)(s')$ for all $s' \neq s$, so it remains to show the equality for the sth coordinates. We have the desired equality

$$\widehat{\mathcal{T}}^{(s,u^\star(s))}_\gamma \left(\widehat{V}^{\star}_{\gamma,M} \right)(s) = u^\star(s) + \gamma \widehat{V}^{\star}_{\gamma,M}(s) = \widehat{\mathcal{T}}_\gamma \left(\widehat{V}^{\star}_{\gamma,M} \right)$$

using the definition of $\widehat{\mathcal{T}}^{(s,u^\star(s))}_\gamma$ and then the definition of $u^\star(s)$, so (67) holds and thus $\widehat{V}^{(s,u^\star(s))}_{\gamma,M} = \widehat{V}^{\star}_{\gamma,M}$.

Next we check that $u^\star(s) \in [0,1]$. We consider two cases, either $\widehat{\mathcal{T}}_\gamma(\widehat{V}^\star_{\gamma,M})(s) = \widehat{V}^\star_{\gamma,M}(s)$ (the clipping operator does not affect entry s) or $\widehat{\mathcal{T}}_\gamma(\widehat{V}^\star_{\gamma,M})(s) > \widehat{V}^\star_{\gamma,M}(s)$. In the first case we immediately have

$$u^\star(s) = \widehat{\mathcal{T}}_\gamma(\widehat{V}^\star_{\gamma,M})(s) - \gamma \widehat{V}^\star_{\gamma,M}(s) = (1-\gamma)\widehat{V}^\star_{\gamma,M}(s)$$

which is $\in [0,1]$ since $\widehat{V}^\star_{\gamma,M}(s) \in [0, \frac{1}{1-\gamma}]$. (This fact can be seen by noting that $\widehat{\mathcal{T}}_\gamma(x) \geq (\text{Clip}_M \circ \widehat{\mathcal{T}}_\gamma)(x)$ for any x so by a monotonicity argument we have that $\widehat{V}^\star_{\gamma,M} \leq \widehat{V}^\star_\gamma \leq \frac{1}{1-\gamma}\mathbf{1}$.) In the second case, since the clipping affects entry s, this means that $\widehat{V}^\star_{\gamma,M}(s)$ must be the largest (not necessarily the uniquely largest) entry of $\widehat{V}^\star_{\gamma,M}$. Therefore we have by the definition of $\widehat{\mathcal{T}}_\gamma(\widehat{V}^\star_{\gamma,M})$ that for some $a^\star \in \mathcal{A}$,

$$\widehat{\mathcal{T}}_\gamma(\widehat{V}^\star_{\gamma,M})(s) = r(s,a^\star) + \gamma \widehat{P}_{s,a^\star}\widehat{V}^\star_{\gamma,M} \leq r(s,a^\star) + \gamma \widehat{V}^\star_{\gamma,M}(s)$$

since $\widehat{P}_{s,a^\star}\widehat{V}^\star_{\gamma,M}$ is an average of entries of $\widehat{V}^\star_{\gamma,M}$ which must be less than the largest entry $\widehat{V}^\star_{\gamma,M}(s)$. After rearranging we have that

$$u^\star(s) = \widehat{\mathcal{T}}_\gamma(\widehat{V}^\star_{\gamma,M})(s) - \gamma \widehat{V}^\star_{\gamma,M}(s) \leq r(s,a^\star) \leq 1,$$

and to lower bound $u^\star(s)$ we can simply use that in this case we have assumed $\widehat{\mathcal{T}}_\gamma(\widehat{V}^\star_{\gamma,M})(s) > \widehat{V}^\star_{\gamma,M}(s)$, so

$$u^\star(s) = \widehat{\mathcal{T}}_\gamma(\widehat{V}^\star_{\gamma,M})(s) - \gamma \widehat{V}^\star_{\gamma,M}(s) > (1-\gamma)\widehat{V}^\star_{\gamma,M}(s) \geq 0.$$

Now we show the penultimate statement. Letting U be a set of $\left\lceil \frac{1}{2(1-\gamma)\varepsilon} \right\rceil$ equally spaced points in $[0,1]$, for any $z \in [0,1]$ there exists $u \in U$ such that $|x-u| \leq (1-\gamma)\varepsilon$. Therefore for any s, letting $U(s)$ be the closest element of U to $u^\star(s)$, we have that

$$\left\|\widehat{V}^\star_{\gamma,M} - \widehat{V}^{(s,U(s))}_{\gamma,M}\right\|_\infty = \left\|\widehat{V}^{(s,u^\star(s))}_{\gamma,M} - \widehat{V}^{(s,U(s))}_{\gamma,M}\right\|_\infty \leq \frac{|u^\star(s) - U(s)|}{1-\gamma} \leq \varepsilon$$

as desired.

Finally, for any $s \in \mathcal{S}, a \in \mathcal{A}, u \in [0,1]$, the independence of $\widehat{V}^{(s,u)}_{\gamma,M}$ from the samples $S^1_{s,a}, \ldots, S^n_{s,a}$ is immediate from the construction, since $\widehat{V}^{(s,u)}_{\gamma,M}$ uses the transition kernel $\widehat{P}^{(s)}$ which is independent of $S^1_{s,a}, \ldots, S^n_{s,a}$. ■

Now we can use this construction to prove the desired error bound. First we state the desired bound:

Lemma 21 *Fix $\gamma \in (0,1)$ and $\delta, M, n > 0$. Let $\widehat{\pi}, \widetilde{V}, \tilde{r}$ be the output of the Span-Constrained Planning Algorithm 3 with inputs $(\widehat{P}, r, \gamma), M, \frac{1}{n}$. Then with probability at least $1-\delta$, we have that*

$$\left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - \widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}\right\|_\infty \leq \frac{24\log_2\log_2(\frac{1}{1-\gamma}+4)}{1-\gamma}\sqrt{\frac{\left\|\widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}\right\|_{\text{span}}+1}{n}16\log\left(\frac{12SAn}{(1-\gamma)^2\delta}\right)}.$$

At a high level, to prove Lemma 21, we take advantage of the formal similarity between our absorbing MDP construction, designed for span-constrained MDPs, and the original construction of Agarwal et al. (2020). Specifically, we are able to reuse certain proof steps from Zurek and Chen (2024) which utilize the absorbing MDP construction of Agarwal et al. (2020), since although their statements involve different objects, their proofs utilize certain properties of said objects which all hold in an identical manner for the objects we consider. First, by examining the specific properties used within the proof of (Zurek and Chen, 2024, Lemma 20), the following result is actually shown:

Lemma 22 *Suppose there exists a random variable $\widehat{V}^\star$, a set U, as well as random variables $(\widehat{V}^\star_{s,u})_{s\in\mathcal{S},u\in U}$ such that*

1. *For any $s\in\mathcal{S}$, $u\in U$, and $a\in\mathcal{A}$, $\widehat{V}^\star_{s,u}$ is independent of the samples $S^1_{s,a},\dots,S^n_{s,a}$ used to construct $\widehat{P}(\cdot\mid s,a)$.*
2. *Almost surely, for any $s\in\mathcal{S}$, there exists $u(s)\in U$ such that $\left\|\widehat{V}^\star-\widehat{V}^\star_{s,u(s)}\right\|_\infty\le\frac{1}{n}$.*
3. *$|U|\le\left\lceil\frac{n}{2(1-\gamma)}\right\rceil$.*

Also suppose $n\ge 4$. With probability at least $1-\delta$, the following holds: for all policies π and all $X\in\mathbb{R}^{\mathcal{S}}$ which satisfy $\left\|X-\widehat{V}^\star\right\|_\infty\le\frac{1}{n}$ and that $\|X\|_{\text{span}}\le\frac{1}{1-\gamma}$, letting $\overline{V}=X-(\min_s X(s))\,\mathbf{1}$, for all $k=0,\dots,\left\lceil\log_2\log_2\left(\|X\|_{\text{span}}+4\right)\right\rceil$, we have

$$\left|\left(\widehat{P}_\pi-P_\pi\right)\left(\overline{V}\right)^{\circ 2^k}\right|\le\sqrt{\frac{\beta\mathbb{V}_{\widehat{P}_\pi}\left[\left(\overline{V}\right)^{\circ 2^k}\right]}{n}}+\frac{\beta\cdot 2^k}{n}\left(\left\|\overline{V}\right\|_\infty+1\right)^{2^k}\mathbf{1}\tag{68}$$

where $\beta=16\log\left(12\frac{SAn}{(1-\gamma)^2\delta}\right)$.

Here $x^{\circ p}$ denotes the elementwise pth power of any vector x. With Lemma 22, as well as another key result from Zurek and Chen (2024) which utilizes a conclusion in the form of Lemma 22 to prove an error bound between value functions, we can now prove Lemma 21.

Proof of Lemma 21 By Lemma 20, it is immediate that the assumptions of Lemma 22 are satisfied for $\widehat{V}^\star\leftarrow\widehat{V}^\star_{\gamma,M}$ and $\widehat{V}^\star_{s,u}\leftarrow\widehat{V}^{(s,u)}_{\gamma,M}$. We would like to obtain the conclusion (68) for $X\leftarrow\widehat{V}^{\widehat{\pi}}_{\gamma,M}$, which we now argue satisfies the assumptions. First, by Lemma 3, we have that (almost surely) $\left\|\widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}-\widehat{V}^\star_{\gamma,M}\right\|_\infty\le\frac{1}{n}$. Second, we have $\mathbf{0}\le\widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}\le\frac{1}{1-\gamma}\mathbf{1}$, which implies $\left\|\widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}\right\|_{\text{span}}\le\frac{1}{1-\gamma}$. Therefore, assuming that $n\ge 4$, by Lemma 22 we have with probability at least $1-\delta$ that, letting $\overline{V}=\widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}-\left(\min_s\widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}(s)\right)\mathbf{1}$ and $\ell=\left\lceil\log_2\log_2\left(\left\|\widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}\right\|_{\text{span}}+4\right)\right\rceil$, for all $k=0,\dots,\ell$, we have

$$\left|\left(\widehat{P}_\pi-P_\pi\right)\left(\overline{V}\right)^{\circ 2^k}\right|\le\sqrt{\frac{\beta\mathbb{V}_{\widehat{P}_\pi}\left[\left(\overline{V}\right)^{\circ 2^k}\right]}{n}}+\frac{\beta\cdot 2^k}{n}\left(\left\|\overline{V}\right\|_\infty+1\right)^{2^k}\mathbf{1}\tag{69}$$

for $\beta=16\log\left(12\frac{SAn}{(1-\gamma)^2\delta}\right)$.

Now we can immediately apply (69) with (Zurek and Chen, 2024, Lemma 16) to obtain that

$$\left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - \widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}\right\|_\infty \leq \frac{4(\ell+1)\beta}{(1-\gamma)n}\left\|\widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}\right\|_{\text{span}} + \frac{2(\ell+1)}{1-\gamma}\sqrt{\frac{2\beta(\left\|\widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}\right\|_{\text{span}}+1)}{n}}. \tag{70}$$

Now all that remains is to simplify (70) which can be done in an identical manner as to (Zurek and Chen, 2024, Proof of Theorem 9) to obtain that

$$\left\|V^{\widehat{\pi}}_{\gamma,\tilde{r}} - \widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}\right\|_\infty \leq \frac{24\log_2\log_2(\frac{1}{1-\gamma}+4)}{1-\gamma}\sqrt{\frac{\left\|\widehat{V}^{\widehat{\pi}}_{\gamma,\tilde{r}}\right\|_{\text{span}}+1}{n}16\log\left(\frac{12SAn}{(1-\gamma)^2\delta}\right)}.$$

■

Now we can apply these error bounds to the analysis of Algorithm 4. We recall the definitions of some objects which appear in Algorithm 4 for convenience, which will be in effect for the remainder of this subsection. We define the empirical transition kernel $\widehat{P}(s' \mid s,a) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{S^i_{s,a} = s'\}$, for all $s' \in \mathcal{S}$, using the n samples drawn from all state-action pairs within Algorithm 4. For all integers $i \in \{2,\ldots,\lceil\log_2 n\rceil$ we define $M_i = 2^i$, and we define γ_i so that $\frac{1}{1-\gamma_i} = \max\left\{\frac{\sqrt{nM_i}}{\alpha(\delta,n)2\sqrt{2}}, 1\right\}$. For each $i \in \{2,\ldots,\lceil\log_2 n\rceil$ we also define $\widetilde{\pi}_i$, $\widetilde{V}_i$, and $\tilde{r}_i$ as the outputs from the Span-Constrained Planning Algorithm 3 with input DMDP $(\widehat{P}, r, \gamma_i)$, input span constraint bound M_i, and input target error $\frac{1}{n}$. For the remainder of this proof we fix a policy π such that ρ^π is a constant vector. (π will later be chosen to optimize a certain deterministic quantity.)

Lemma 23 *Define the function* $\alpha(\tilde{\delta},\tilde{n}) = 24\sqrt{16\log\left(\frac{24SA\tilde{n}^5}{\tilde{\delta}}\right)}\log_2(\log_2(\tilde{n}+4))$. *Fix* $\delta > 0$. *Then with probability at least* $1-\delta$, *we have for all* $i = 2,\ldots,\lceil\log_2 n\rceil$ *that*

$$\left\|V^\pi_{\gamma_i} - \widehat{V}^\pi_{\gamma_i}\right\|_\infty \leq \frac{\alpha(\delta,n)}{1-\gamma_i}\sqrt{\frac{\left\|V^\pi_{\gamma_i}\right\|_{\text{span}}+1}{n}} \tag{71}$$

and also the Span-Constrained Planning Algorithm 3 used within subroutine on line 8 outputs a policy $\widetilde{\pi}_i$, *approximate value function* $\widetilde{V}_i$, *and reward function* $\tilde{r}_i$ *such that*

$$\left\|\widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i} - \widehat{V}^\star_{\gamma_i,M_i}\right\|_\infty \leq \frac{1}{n} \tag{72}$$

$$\left\|\widetilde{V}_i - \widehat{V}^\star_{\gamma_i,M_i}\right\|_\infty \leq \frac{1}{n} \tag{73}$$

$$\left\|\widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i} - V^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_\infty \leq \frac{\alpha(\delta,n)}{1-\gamma_i}\sqrt{\frac{\left\|\widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_{\text{span}}+1}{n}}. \tag{74}$$

Proof First, we note that the properties (72) and (73) immediately follow from Lemma 3. By a union bound and Lemmas 19 and 21, we have that with probability at least $1-2(\lceil\log_2 n\rceil - 1)\delta'$, for all $i = 2,\ldots,\lceil\log_2 n\rceil$, it holds that

$$\left\|V^\pi_{\gamma_i} - \widehat{V}^\pi_{\gamma_i}\right\|_\infty \leq \frac{24\log_2\log_2(\frac{1}{1-\gamma_i}+4)}{1-\gamma_i}\sqrt{\frac{\left\|V^\pi_{\gamma_i}\right\|_{\text{span}}+1}{n}16\log\left(\frac{12SAn}{(1-\gamma_i)^2\delta'}\right)}.$$

and

$$\left\|V^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i} - \widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_\infty \le \frac{24\log_2\log_2(\frac{1}{1-\gamma_i}+4)}{1-\gamma_i}\sqrt{\frac{\left\|\widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_{\text{span}}+1}{n}16\log\left(\frac{12SAn}{(1-\gamma_i)^2\delta'}\right)}$$

Since

$$\frac{1}{1-\gamma_i} \le \frac{1}{1-\gamma_{\lceil\log_2 n\rceil}} \le \max\left\{1, \frac{\sqrt{nM_{\lceil\log_2 n\rceil}}}{\alpha(\delta,n)2\sqrt{2}}\right\} \le \max\left\{1, \frac{\sqrt{n2n}}{\alpha(\delta,n)2\sqrt{2}}\right\} \le n$$

since $\alpha(\delta,n) \ge 1$ and $n \ge 1$, and also noting that $\lceil\log_2 n\rceil - 1 \le \log_2 n \le n$, by taking $\delta' = \frac{\delta}{2(\lceil\log_2 n\rceil-1)}$ we obtain that with probability at least $1-\delta$ both

$$\begin{aligned}\left\|V^{\pi}_{\gamma_i} - \widehat{V}^{\pi}_{\gamma_i}\right\|_\infty &\le \frac{24\log_2\log_2(n+4)}{1-\gamma_i}\sqrt{\frac{\left\|V^{\pi}_{\gamma_i}\right\|_{\text{span}}+1}{n}16\log\left(\frac{24SAn^3(\lceil\log_2 n\rceil-1)}{\delta}\right)} \\ &\le \frac{\alpha(\delta,n)}{1-\gamma_i}\sqrt{\frac{\left\|V^{\pi}_{\gamma_i}\right\|_{\text{span}}+1}{n}}\end{aligned}$$

and similarly

$$\left\|\widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i} - V^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_\infty \le \frac{\alpha(\delta,n)}{1-\gamma_i}\sqrt{\frac{\left\|\widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_{\text{span}}+1}{n}}$$

for all $i = 2,\dots,\lceil\log_2 n\rceil$. ■

Lemma 24 *Under the event described in Lemma 23, we have*

$$\rho^{\widetilde{\pi}_i} \ge \widehat{L}(i)\mathbf{1} = (1-\gamma_i)\min_s \widetilde{V}_i(s)\mathbf{1} - 2\frac{1-\gamma_i}{n}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{M_i+\frac{2}{n}+1}{n}}\mathbf{1} \tag{75}$$

for all $i = 2,\dots,\lceil\log_2 n\rceil$.

Proof The proof is very similar to the first part of the proof of Lemma 16. Fix $i \in \{2,\dots,\lceil\log_2 n\rceil\}$. First, by using (72) and the fact that $\left\|\widehat{V}^{\star}_{\gamma_i,M_i}\right\|_{\text{span}} \le M_i$, we have

$$\left\|\widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_{\text{span}} \le \left\|\widehat{V}^{\star}_{\gamma_i,M_i}\right\|_{\text{span}} + \left\|\widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i} - \widehat{V}^{\star}_{\gamma_i,M_i}\right\|_{\text{span}} \le \left\|\widehat{V}^{\star}_{\gamma_i,M_i}\right\|_{\text{span}} + 2\left\|\widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i} - \widehat{V}^{\star}_{\gamma_i,M_i}\right\|_\infty \le M_i + \frac{2}{n}. \tag{76}$$

We note that by triangle inequality and (72) and (73) we have $\left\|\widetilde{V}_i - \widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_\infty \le \frac{2}{n}$. Using this bound on $\left\|\widetilde{V}_i - \widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_\infty$, (74), and (76), we have

$$\begin{aligned}
\rho^{\widetilde{\pi}_i} &\ge (1-\gamma_i)\min_s V^{\widetilde{\pi}_i}_{\gamma_i}(s)\mathbf{1} \\
&\ge (1-\gamma_i)\min_s V^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}(s)\mathbf{1} \\
&\ge (1-\gamma_i)\min_s \widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}(s)\mathbf{1} - (1-\gamma_i)\left\|V^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i} - \widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_\infty \mathbf{1} \\
&\ge (1-\gamma_i)\min_s \widetilde{V}_i(s)\mathbf{1} - (1-\gamma_i)\left\|\widetilde{V}_i - \widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_\infty \mathbf{1} - (1-\gamma_i)\left\|V^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i} - \widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_\infty \mathbf{1} \\
&\ge (1-\gamma_i)\min_s \widetilde{V}_i(s)\mathbf{1} - 2\frac{1-\gamma_i}{n}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{\left\|\widehat{V}^{\widetilde{\pi}_i}_{\gamma_i,\tilde{r}_i}\right\|_{\text{span}} + 1}{n}}\mathbf{1} \\
&\ge (1-\gamma_i)\min_s \widetilde{V}_i(s)\mathbf{1} - 2\frac{1-\gamma_i}{n}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{M_i + \frac{2}{n} + 1}{n}}\mathbf{1} \\
&= \widehat{L}(i)\mathbf{1}.
\end{aligned}$$

■

Before continuing we need to establish some relationships between DMDP and AMDP quantities for the policy π. This lemma is similar to (Wei et al., 2020, Lemma 2) but for a general policy.

Lemma 25 *Suppose ρ^π is constant. Then*

1. $\left\|\frac{1}{1-\gamma}\rho^\pi - V^\pi_\gamma\right\|_\infty \le \|h^\pi\|_{\text{span}}$
2. $\left\|V^\pi_\gamma\right\|_{\text{span}} \le 2\|h^\pi\|_{\text{span}}$.

Proof For the first statement, we have that

$$V^\pi_\gamma = (I-\gamma P_\pi)^{-1} r_\pi = (I-\gamma P_\pi)^{-1}(\rho^\pi + h^\pi - P_\pi h^\pi) = (I-\gamma P_\pi)^{-1}\rho^\pi + (I-\gamma P_\pi)^{-1}(I-P_\pi)h^\pi.$$

Since $P_\pi \rho^\pi = \rho^\pi$, by the Neumann series the first term is equal to $\frac{1}{1-\gamma}\rho^\pi$. By a standard calculation (e.g. (Zurek and Chen, 2025, Lemma 20)) the second term satisfies $\left\|(I-\gamma P_\pi)^{-1}(I-P_\pi)h^\pi\right\|_\infty \le \|h^\pi\|_{\text{span}}$. Therefore we have that $\left\|\frac{1}{1-\gamma}\rho^\pi - V^\pi_\gamma\right\|_\infty \le \left\|(I-\gamma P_\pi)^{-1}(I-P_\pi)h^\pi\right\|_\infty \le \|h^\pi\|_{\text{span}}$.

For the second statement, since ρ^π is constant, we have that

$$\left\|V^\pi_\gamma\right\|_{\text{span}} = \left\|V^\pi_\gamma - \frac{1}{1-\gamma}\rho^\pi\right\|_{\text{span}} \le 2\left\|V^\pi_\gamma - \frac{1}{1-\gamma}\rho^\pi\right\|_\infty \le 2\|h^\pi\|_{\text{span}}.$$

■

Lemma 26 *Suppose that for some integer $i \ge 2$ we have $\|h^\pi\|_{\text{span}} + 1 \le M_i/4$, and also that $n \ge 2\alpha(\delta,n)^2$. Then under the event in Lemma 23, we have*

$$\widehat{L}(i)\mathbf{1} \ge \rho^\pi - \alpha(\delta,n)(2+\sqrt{2})\sqrt{\frac{M_i}{n}}\mathbf{1}.$$

Proof The fact that $n \geq 2\alpha(\delta,n)^2$ implies that we always have

$$\frac{1}{1-\gamma_i} = \max\left\{\frac{\sqrt{nM_i}}{\alpha(\delta,n)2\sqrt{2}}, 1\right\} = \frac{\sqrt{nM_i}}{\alpha(\delta,n)2\sqrt{2}}$$

for all $i \geq 2$, since $M_i = 2^i \geq 4$. Now, by triangle inequality, $\|\cdot\|_{\text{span}} \leq 2\|\cdot\|_\infty$, (71), Lemma 25, and the above expression for $\frac{1}{1-\gamma_i}$, we calculate that

$$\begin{aligned}
\left\|\widehat{V}^\pi_{\gamma_i}\right\|_{\text{span}} &\leq \left\|V^\pi_{\gamma_i}\right\|_{\text{span}} + \left\|\widehat{V}^\pi_{\gamma_i} - V^\pi_{\gamma_i}\right\|_{\text{span}} \\
&\leq \left\|V^\pi_{\gamma_i}\right\|_{\text{span}} + 2\left\|\widehat{V}^\pi_{\gamma_i} - V^\pi_{\gamma_i}\right\|_\infty \\
&\leq \left\|V^\pi_{\gamma_i}\right\|_{\text{span}} + 2\frac{\alpha(\delta,n)}{1-\gamma_i}\sqrt{\frac{\left\|V^\pi_{\gamma_i}\right\|_{\text{span}} + 1}{n}} \\
&\leq 2\left\|h^\pi\right\|_{\text{span}} + 2\frac{\alpha(\delta,n)}{1-\gamma_i}\sqrt{\frac{2\left\|h^\pi\right\|_{\text{span}} + 1}{n}} \\
&= 2\left\|h^\pi\right\|_{\text{span}} + \frac{\sqrt{nM_i}}{\sqrt{2}}\sqrt{\frac{2\left\|h^\pi\right\|_{\text{span}} + 1}{n}} \\
&\leq \frac{M_i}{2} + \frac{\sqrt{M_i}}{\sqrt{2}}\sqrt{M_i/2} \\
&= M_i.
\end{aligned}$$

Consequently by Lemma 3 we have that $\widehat{V}^\star_{\gamma_i,M_i} \geq \widehat{V}^\pi_{\gamma_i}$, and thus

$$\widetilde{V}_i \geq \widehat{V}^\star_{\gamma_i,M_i} - \frac{1}{n}\mathbf{1} \geq \widehat{V}^\pi_{\gamma_i} - \frac{1}{n}\mathbf{1}.$$

Now we lower-bound $\widehat{V}^\pi_{\gamma_i}$. We have that

$$\begin{aligned}
\widehat{V}^\pi_{\gamma_i} &\geq V^\pi_{\gamma_i} - \left\|\widehat{V}^\pi_{\gamma_i} - V^\pi_{\gamma_i}\right\|_\infty \mathbf{1} \\
&\geq \frac{1}{1-\gamma_i}\rho^\pi - \left\|\frac{1}{1-\gamma_i}\rho^\pi - V^\pi_{\gamma_i}\right\|_\infty \mathbf{1} - \left\|\widehat{V}^\pi_{\gamma_i} - V^\pi_{\gamma_i}\right\|_\infty \mathbf{1} \\
&\geq \frac{1}{1-\gamma_i}\rho^\pi - \left\|h^\pi\right\|_{\text{span}}\mathbf{1} - \frac{\alpha(\delta,n)}{1-\gamma_i}\sqrt{\frac{\left\|V^\pi_{\gamma_i}\right\|_{\text{span}} + 1}{n}}\mathbf{1} \qquad (77)
\end{aligned}$$

using Lemma 25 and (71) again. Thus

$$\begin{aligned}
\widetilde{V}_i &\geq \widehat{V}^\pi_{\gamma_i} - \frac{1}{n}\mathbf{1} \\
&\geq \frac{1}{1-\gamma_i}\rho^\pi - \left\|h^\pi\right\|_{\text{span}}\mathbf{1} - \frac{\alpha(\delta,n)}{1-\gamma_i}\sqrt{\frac{\left\|V^\pi_{\gamma_i}\right\|_{\text{span}} + 1}{n}}\mathbf{1} - \frac{1}{n}\mathbf{1} \\
&\geq \frac{1}{1-\gamma_i}\rho^\pi - \frac{M_i}{4}\mathbf{1} - \frac{\alpha(\delta,n)}{1-\gamma_i}\sqrt{\frac{2\left\|h^\pi\right\|_{\text{span}} + 1}{n}} \\
&\geq \frac{1}{1-\gamma_i}\rho^\pi - \frac{M_i}{4}\mathbf{1} - \frac{\alpha(\delta,n)}{1-\gamma_i}\sqrt{\frac{M_i/2}{n}}
\end{aligned}$$

where in the first inequality we combine (73) with Lemma 3, which states that $\widehat{V}^{\star}_{\gamma_i,M_i} \geq \widehat{V}^{\pi}_{\gamma_i}$ since (as we calculated above) we have $\left\|\widehat{V}^{\pi}_{\gamma_i}\right\|_{\text{span}} \leq M_i$. In the second inequality we use (77), in the third we use Lemma 25, and in the final inequality we use the assumption $\|h^{\pi}\|_{\text{span}} + 1 \leq M_i/4$. Therefore, we have

$$\begin{aligned}
\widehat{L}(i) &= (1-\gamma_i)\min_s \widetilde{V}_i(s) - 2\frac{1-\gamma_i}{n} - \alpha(\delta,n)\sqrt{\frac{M_i + \frac{2}{n} + 1}{n}} \\
&\geq \rho^{\pi}(s_0) - (1-\gamma_i)\frac{M_i}{4}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{M_i/2}{n}} - 2\frac{1-\gamma_i}{n} - \alpha(\delta,n)\sqrt{\frac{M_i + \frac{2}{n} + 1}{n}} \\
&= \rho^{\pi}(s_0) - \frac{\alpha(\delta,n)2\sqrt{2}}{\sqrt{nM_i}}\frac{M_i}{4}\mathbf{1} - \alpha(\delta,n)\sqrt{\frac{M_i/2}{n}} - 2\frac{1-\gamma_i}{n} - \alpha(\delta,n)\sqrt{\frac{M_i + \frac{2}{n} + 1}{n}} \\
&\geq \rho^{\pi}(s_0) - \alpha(\delta,n)\sqrt{2}\sqrt{\frac{M_i}{n}} - \frac{1}{\sqrt{2}}\frac{1}{n}\frac{1}{2\sqrt{n}} - \alpha(\delta,n)\sqrt{\frac{M_i + \frac{2}{n} + 1}{n}} \\
&\geq \rho^{\pi}(s_0) - \alpha(\delta,n)(2+\sqrt{2})\sqrt{\frac{M_i}{n}},
\end{aligned}$$

where s_0 is an arbitrary state (since ρ^{π} is constant), and we used that $\alpha(\delta,n) \geq 1$ and $M_i \geq 4$. ■

Proof of Theorem 4 We assume that the event in Lemma 23 holds, which occurs with probability at least $1-\delta$. Lemma 24 immediately implies that

$$\rho^{\widehat{\pi}} = \rho^{\widetilde{\pi}_{\widehat{i}}} \geq \widehat{L}(\widehat{i})\mathbf{1}.$$

We also trivially have that $\rho^{\widehat{\pi}} \geq \mathbf{0}$, which implies that

$$\rho^{\widehat{\pi}} \geq \max\{\widehat{L}(\widehat{i}), 0\}\mathbf{1} = \widehat{\rho}.$$

Now we lower-bound $\widehat{\rho}$. First note that if $n < 2\alpha(\delta,n)^2$, then

$$10\alpha(\delta,n)\sqrt{\frac{\|h^{\pi}\|_{\text{span}} + 1}{n}} \geq 10/\sqrt{2} > 1,$$

so the desired conclusion holds trivially since we must have $\widehat{\rho} \geq \mathbf{0}$, and $\rho^{\pi} \leq \mathbf{1}$. Thus we can now consider the situation $n \geq 2\alpha(\delta,n)^2$. We note that the smallest $i = 2$ causes $M_2/8 = \frac{1}{2} < 1$. Therefore by the construction of the M_i's, either there exists i such that

$$M_i/8 \leq \|h^{\pi}\|_{\text{span}} + 1 \leq M_i/4 \tag{78}$$

or we have that $\|h^{\pi}\|_{\text{span}} + 1 > M_i/4$ for all i. Since the largest $i = \lceil \log_2 n \rceil$ causes $M_i/4 > n/4$, in this second case we thus have that

$$10\alpha(\delta,n)\sqrt{\frac{\|h^{\pi}\|_{\text{span}} + 1}{n}} \geq 10\alpha(\delta,n)\frac{1}{2} > 1,$$

so again the desired conclusion holds trivially since we must have $\widehat{\rho} \geq \mathbf{0}$. In the first case that (78) holds for some i, since also $n \geq 2\alpha(\delta, n)^2$, by Lemma 26 we have that

$$\begin{aligned}\widehat{\rho} \geq \widehat{L}(\widehat{i})\mathbf{1} \geq \widehat{L}(i)\mathbf{1} &\geq \rho^\pi - \alpha(\delta, n)(2+\sqrt{2})\sqrt{\frac{M_i}{n}}\mathbf{1} \\ &\geq \rho^\pi - (2+\sqrt{2})\sqrt{8}\alpha(\delta, n)\sqrt{\frac{\|h^\pi\|_{\text{span}} + 1}{n}} \\ &\geq \rho^\pi - 10\alpha(\delta, n)\sqrt{\frac{\|h^\pi\|_{\text{span}} + 1}{n}}.\end{aligned}$$

Since we have proven this for arbitrary fixed π, we can apply this result to a policy $\pi \in \sup_{\pi:\rho^\pi \text{ constant}} \rho^\pi(s_0) - 10\alpha(\delta, n)\sqrt{\frac{\|h^\pi\|_{\text{span}}+1}{n}}$ (for an arbitrary state s_0) to obtain the desired conclusion. Finally, we can set $C_4 = 10$. ■

Appendix D. Examples

In this section we provide the two examples mentioned in Subsection 3.3 of situations where the guarantee of Theorem 4 could be much better than the minimax rate. In both examples each solid line represents a single action, and an expression "$R = \ldots$" denotes the reward for said action. If the line splits into multiple dashed arrows then this indicates that the next-state transition from following this action is stochastic, and the numbers next to each dashed line indicate the transition probabilities. Otherwise if the line does not split then it indicates a deterministic next-state transition.

First we provide the simpler of the two examples, where $\|h^\star\|_{\text{span}}$ is arbitrarily larger than $\inf_{\pi:\rho^\pi=\rho^\star} \|h^\pi\|_{\text{span}}$.

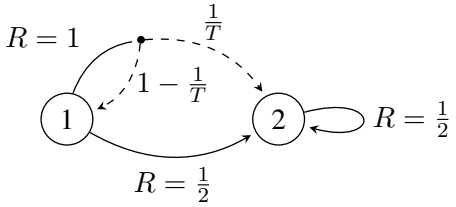

Figure 1: An MDP where $\|h^\star\|_{\text{span}}$ can be arbitrarily larger than $\inf_{\pi:\rho^\pi=\rho^\star} \|h^\pi\|_{\text{span}}$.

Theorem 27 *Consider the MDP displayed in Figure 1. For any $T \geq 1$, we have $\|h^\star\|_{\text{span}} = \frac{T}{2}$ and $\inf_{\pi:\rho^\pi=\rho^\star} \|h^\pi\|_{\text{span}} = 0$.*

Next, we provide an instance where $\|h^\star\|_{\text{span}}$ can be arbitrarily large but a policy π with an arbitrarily low level of suboptimality $\rho^\star - \rho^\pi$ can have $\|h^\pi\|_{\text{span}} = O(1)$.

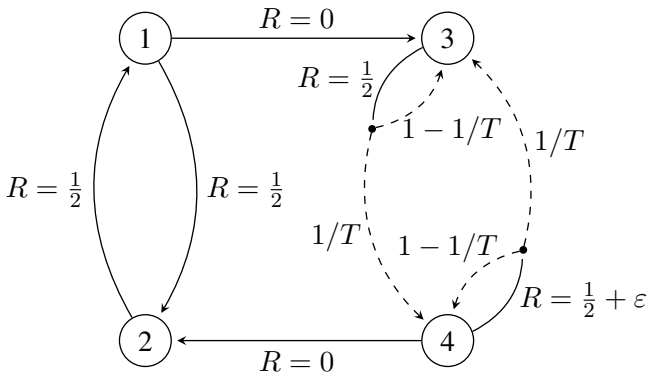

Figure 2: An MDP where $\|h^\star\|_{\text{span}}$ can be arbitrarily larger than $\|h^\pi\|_{\text{span}}$ for some near-optimal policy π satisfying $\rho^\pi = \rho^\star - \frac{\varepsilon}{2}\mathbf{1}$.

Theorem 28 *Consider the MDP displayed in Figure 2. For any $T \geq 1$ and $\varepsilon > 0$, we have that $\|h^\star\|_{\text{span}} = \frac{\varepsilon T}{2} + \varepsilon + \frac{1}{2}$, but there exists some policy π with constant gain such that $\rho^\pi = \rho^\star - \frac{\varepsilon}{2}\mathbf{1}$ and $\|h^\pi\|_{\text{span}} = \frac{1}{2}$.*

D.1. Proofs of Theorems 27 and 28

Proof of Theorem 27 It is easy to see that state 1 is transient under all policies and state 2 is absorbing, so all policies are gain-optimal and have gain $\frac{1}{2}\mathbf{1}$. Only state 1 has multiple possible actions, so it suffices to consider the two policies π_{up}, which takes the "up" action which has nonzero probability of returning to state 1, and π_{down}, which leads to an immediate transition to state 2. It is trivial to see that $h^{\pi_{\text{down}}} = \mathbf{0}$, so we have that $\|h^{\pi_{\text{down}}}\|_{\text{span}} = 0$. To compute $h^{\pi_{\text{up}}}$, we must have $h^{\pi_{\text{up}}}(2) = 0$, so we can then calculate that

$$\begin{aligned} h^{\pi_{\text{up}}}(1) + \frac{1}{2} &= 1 + (1 - \frac{1}{T})h^{\pi_{\text{up}}}(1) + \frac{1}{T}h^{\pi_{\text{up}}}(2) \\ \implies h^{\pi_{\text{up}}}(1) &= \frac{T}{2} + h^{\pi_{\text{up}}}(2) = \frac{T}{2} \end{aligned}$$

and thus $\|h^{\pi_{\text{up}}}\|_{\text{span}} = T/2$. Since these are the only two stationary deterministic policies, one of them must be Blackwell-optimal, and since they have equal gain and elementwise $h^{\pi_{\text{up}}} \geq h^{\pi_{\text{down}}}$ (with a strict inequality in state 1), we have that $\pi_{\text{up}} = \pi^\star$ and $\|h^\star\|_{\text{span}} = \|h^{\pi_{\text{up}}}\|_{\text{span}} = T/2$. Since π_{down} is gain-optimal it is immediate from $h^{\pi_{\text{down}}} = \mathbf{0}$ that $\inf_{\pi:\rho^\pi=\rho^\star} \|h^\pi\|_{\text{span}} = 0$. ■

Proof of Theorem 28 There are two states, 1 and 4, where multiple actions are possible. We name the actions in state 1 the "down" action (which leads to 2) and the "right" action (which leads to 3), and we name the actions in state 4 the "up" action (which has positive probability of leading to 3) and the "left" action (which leads to 2). A deterministic stationary policy can be specified by its two choices between the actions available at states 1 and 4. We thus use π_{DL} to indicate the policy which takes the down action in state 1 and the left action in state 4, and likewise for other choices of $\{D, R\} \times \{U, L\}$.

It is easy to see that since $\varepsilon > 0$ the unique gain-optimal policy is π_{RU} which has $\rho^{\pi_{RU}} = \rho^\star = \frac{1+\varepsilon}{2}\mathbf{1}$. Thus this policy must also be Blackwell-optimal. We now compute $h^\star = h^{\pi_{RU}}$. We have

$$\begin{aligned} h^\star(3) + \frac{1+\varepsilon}{2} &= \frac{1}{2} + (1 - \frac{1}{T})h^\star(3) + \frac{1}{T}h^\star(4) \\ \implies h^\star(3) &= h^\star(4) - \frac{\varepsilon T}{2} \end{aligned}$$

and since the stationary distribution of π_{RU} has equal mass on states 3 and 4 we must have $h^\star(3) + h^\star(4) = 0$, which implies that $h^\star(3) = -\frac{\varepsilon T}{4}$ and $h^\star(4) = \frac{\varepsilon T}{4}$. It is then easy to check that $h^\star(1) = h^\star(3) - \frac{1+\varepsilon}{2} = \frac{-\varepsilon T - 2\varepsilon - 2}{4}$ and $h^\star(2) = h^\star(1) - \frac{\varepsilon}{2} = \frac{-\varepsilon T - 4\varepsilon - 2}{4}$, so we have that $\|h^\star\|_{\text{span}} = \frac{\varepsilon T}{4} - \frac{-\varepsilon T - 4\varepsilon - 2}{4} = \frac{\varepsilon T}{2} + \varepsilon + \frac{1}{2}$.

Next we consider the policy π_{DL}. It is easy to see that $\rho^{\pi_{DL}} = \frac{1}{2}\mathbf{1}$, which is constant and which satisfies $\rho^\star - \rho^{\pi_{DL}} = \frac{\varepsilon}{2}\mathbf{1}$. Now we compute $\|h^{\pi_{DL}}\|_{\text{span}}$. It is immediate to see that $h^{\pi_{DL}}(1) = h^{\pi_{DL}}(2) = 0$. Then we can calculate that $h^{\pi_{DL}}(4) = h^{\pi_{DL}}(2) - \frac{1}{2} = -\frac{1}{2}$, and finally that

$$\begin{aligned} h^{\pi_{DL}}(3) + \frac{1}{2} &= \frac{1}{2} + (1 - \frac{1}{T})h^{\pi_{DL}}(3) + \frac{1}{T}h^{\pi_{DL}}(4) \\ \implies h^{\pi_{DL}}(3) &= h^{\pi_{DL}}(4) = -\frac{1}{2}. \end{aligned}$$

Therefore we have that $\|h^{\pi_{DL}}\|_{\text{span}} = 0 - -\frac{1}{2} = \frac{1}{2}$. ■