Almost Horizon-Free Structure-Aware Best Policy Identification with a Generative Model

Andrea Zanette
Institute for Computational and Mathematical Engineering,
Stanford University, CA
zanette@stanford.edu

Mykel J. Kochenderfer
Department of Aeronautics and Astronautics,
Stanford University, CA
mykel@stanford.edu

Emma Brunskill
Department of Computer Science,
Stanford University, CA
ebrun@cs.stanford.edu

Abstract

This paper focuses on the problem of computing an ϵ-optimal policy in a discounted Markov Decision Process (MDP) provided that we can access the reward and transition function through a generative model. We propose an algorithm that is initially agnostic to the MDP but that can leverage the specific MDP structure, expressed in terms of variances of the rewards and next-state value function, and gaps in the optimal action-value function to reduce the sample complexity needed to find a good policy, precisely highlighting the contribution of each state-action pair to the final sample complexity. A key feature of our analysis is that it removes all horizon dependencies in the sample complexity of suboptimal actions except for the intrinsic scaling of the value function and a constant additive term.

1 Introduction

A key goal is to design reinforcement learning (RL) agents that can leverage problem structure to efficiently learn a good policy, especially in problems with very long time horizons. Ideally the RL algorithm should be able to adjust without apriori information about the problem structure. Formal analyses that characterize the performance of such algorithms yielding instance-dependent bound help to advance our core understanding of the characteristics that govern the hardness of learning to make good decisions under uncertainty.

Though there is relatively limited work in reinforcement learning, strong problem-dependent guarantees are available for multi-armed bandits. In particular, well known bounds for online learning scale as a function of the gap between the expected reward of a particular action and the optimal action [ABF02] and also on the variance of the rewards [AMS09]. In the pure exploration setting in bandits, which is related to the setting we consider in this paper, there exist multiple algorithms with problem-dependent bounds [EMM06; MM94; MSA08; Jam+14; BMS09; ABM10; GGL12; KKS13] of this form. Ideally the complexity of learning to make good decisions in reinforcement learning tasks would scale with previously identified quantities of gap and variance over the next value function. As a step towards this, in this paper we introduce an algorithm for an RL agent operating in a discrete state and action space that has access to a generative model and can leverage problem-dependent structure to have strong instance-dependent PAC sample complexity bounds that are a function of the variance of the rewards and next state value functions, as well as the gaps between the optimal and suboptimal state-action values. While the sequential setting brings additional difficulties due to possibly long horizon, our bounds also show that in the dominant terms, our

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

approach avoids suffering any horizon dependence for suboptimal actions beyond the scaling of the value function. This significantly improves in statistical efficiency over prior worst-case bounds for the generative model case [GMK13; Sid+18] and matches existing worst-case bounds in worst-case settings.

To do so we introduce a novel algorithm structure that acquires samples of state-action pairs in iterative rounds. A slight variant of the well known simulation lemma (see e.g. [KMN02]) suggests that in order to improve our estimate of the optimal value function and policy, it is sufficient to ensure that after each round of sampling, the confidence intervals shrink over the MDP parameter estimates of both the state–action pairs *visited by the optimal policy* and the state–action pairs *visited by the empirically-optimal policy*. While of course both are unknown, we show that we can implicitly maintain a set of candidate policies that are ϵ-accurate, and by ensuring that we shrink the confidence sets of all state–action pairs likely to be visited by any such policy, we are also guaranteed (with high probability) to shrink the confidence intervals of the optimal policy. Interestingly we can show that by focusing on such state–action pairs, we can avoid the horizon dependence on suboptimal actions. The key idea is to take into account the MDP learned dynamics to enforce a constraint on the suboptimality of the candidate policies. The sampling strategy is derived by solving a minimax problem that minimizes the number of samples to guarantee that every policy in the set of candidate policies is accurately estimated. Importantly, this minimax problem can be reformulated as a convex minimization problem that can be solved with any standard solver for convex optimization.

Our algorithmic approach is quite different from many prior approaches, both in the generative model setting and the online setting. When a generative model is available, the available worst-case optimal algorithms [AMK12; Sid+18] allocate samples uniformly to all state and action pairs. We show our approach can be substantialy more effective for general case of MDPs with heterogeneous structure, and even for the pathologically hard instances because of the reduced horizon dependence on suboptimal actions. Note too that our approach is quite different from online RL algorithms that often (implicitly) allocate exploration budget to state-action pairs encountered by the policy with the most optimistic upper bound [JOA10; AOM17; OVR13; DB15; DLB17; SLL09; LH14], since here we explicitly reason about the reduction in the confidence intervals across a large set of policies whose value is near the *empirical* optimal value at this round.

2 Notation and Preliminaries

We consider discounted infinite horizon MDPs [SB18], which are defined by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces with cardinality S and A, respectively. We denote by $p(s' \mid s, a)$ the probability of transitioning to state s' after taking action a in state s while $r(s,a) \in [0,1]$ is the average instantaneous reward collected and $R(s,a) \in [0,1]$ the corresponding random variable. The vector value function of policy π is denoted with V^π. If ρ is the initial starting distribution then $V(\rho) = \sum_s \rho_s V(s)$. The value function of the optimal policy $\pi^\star$ is denoted with $V^\star = V^{\pi^\star}$. We call $\operatorname{Var} R(s,a)$ and $\operatorname{Var}_{p(s,a)} V^\star$ the variance of $R(s,a)$ and of $V^\star(s')$ where $s' \sim p(s,a)$. The agent interacts with the MDP via a generative model that takes as input a (s,a) pair and returns a random sample of the reward $R(s,a)$ and a random next state s^+ according to the transition model $s^+ \sim p(s,a)$. The reinforcement learning agent maintains an empirical MDP $\widehat{\mathcal{M}}_k = \langle \mathcal{S}, \mathcal{A}, \widehat{p}_k, \widehat{r}_k, \gamma \rangle$ for every iteration/episode k, and the maximum likelihood transition $\widehat{p}_k(s,a)$ and rewards $\widehat{r}_k(s,a)$ have received n^k_{sa} samples. The $\widehat{V}^\star_k$ is the empirical estimate using MDP $\widehat{\mathcal{M}}_k$ of the empirical optimal policy $\widehat{\pi}^\star_k$. Variables with a hat refer to the empirical MDP $\widehat{\mathcal{M}}_k$ and the subscript indicates what iteration/episode they refer to. We denote with $\overline{w}^{\pi,\rho}_{sa} = \sum_{t=0}^{\infty} \gamma^t \Pr(s,a,t,\rho)$ the discounted sum of visit probabilities $\Pr(s,a,t,\rho)$ to the (s,a) pair in timestep t if the starting state is drawn uniformly from ρ and $\widehat{\overline{w}}^{\pi,k,\rho}_{sa}$ is its analogous on $\widehat{\mathcal{M}}_k$. We use the $\tilde{O}(\cdot)$ notation to indicate a quantity that depends on $(\cdot)$ up to a polylog expression of a quantity at most polynomial in $S, A, \frac{1}{1-\gamma} \frac{1}{\delta}$, where δ is the "failure probability". Before proceeding, we first recall the following lemma from [GMK13]:

Lemma 2 (Simulation Lemma for Optimal Value Function Estimate [GMK13])**.** *With probability at least $1-\delta$, outside the failure event for any starting distribution ρ it holds that:*

$$(V^\star - \widehat{V}_k^\star)(\rho) \le \sum_{(s,a)} \widehat{\widehat{w}}_{sa}^{\pi^\star,k,\rho}\left((r-\widehat{r}_k)(s,a)+\gamma(p-\widehat{p}_k)(s,a)^\top V^\star\right) \le \sum_{(s,a)} \widehat{\widehat{w}}_{sa}^{\pi^\star,k,\rho} CI_{sa}(n_{sa}^k)$$

$$(V^\star - \widehat{V}_k^\star)(\rho) \ge \sum_{(s,a)} \widehat{\widehat{w}}_{sa}^{\widehat{\pi}_k^\star,k,\rho}\left((r-\widehat{r}_k)(s,a)+\gamma(p-\widehat{p}_k)(s,a)^\top V^\star\right) \ge -\sum_{(s,a)} \widehat{\widehat{w}}_{sa}^{\widehat{\pi}_k^\star,k,\rho} CI_{sa}(n_{sa}^k)$$

The $CI_{sa}(n_{sa}^k)$ are Bernstein's confidence intervals (defined in more details in appendix A) after n_{sa}^k samples over the rewards and transitions and are function of the unknown rewards and transition variances. The proof (see appendix) is a slight variation of lemma 3 in [GMK13].

3 Sampling Strategy Given an Empirical MDP

We first describe how our approach will allocate samples to state–action pairs given a current empirical MDP, before presenting in the next section our full algorithm.

Lemma 1 suggests that to estimate the optimal value function it suffices to accurately estimate the (s,a) pairs in the trajectories identified by two policies, namely the optimal policy $\pi^\star$ (optimal on $\mathcal{M}$) and the empirical optimal policy $\widehat{\pi}_k^\star$ (optimal on $\widehat{\mathcal{M}}_k$). Since $\pi^\star$ and $\widehat{\pi}_k^\star$ are unknown (in particular, $\widehat{\pi}_k^\star$ is a random variable prior to sampling), a common strategy is to allocate an identical number of samples uniformly [GMK13; Sid+18] so that the confidence intervals $CI_{sa}(n_{sa}^k)$ are sufficiently small for all state–action pairs leading to a small $|(V^\star - \widehat{V}_k^\star)(\rho)|$; from here it is possible to show that the empirical optimal policy $\widehat{\pi}_k^\star$ can be extracted and that $|(V^\star - V^{\widehat{\pi}_k^\star})(\rho)|$ is also small (so $\widehat{\pi}_k^\star$ is near-optimal). Therefore, in the main text we mostly focus on showing that $|(V^\star - \widehat{V}_k^\star)(\rho)|$ is small, and defer additional details to the appendix. The proposed approach is to proceed in iterations or episodes. In each episode our algorithm implicitly maintains a set of candidate policies, which are near-optimal, and allocates more samples to the (s,a) pairs visited by these policies to refine their estimated value. On the next episode those policies that are too suboptimal relative to their estimation accuracy are implicitly discarded. In particular, the samples are placed in a way that is related to the visit probabilities of the near-optimal empirical policies in addition to the variances of the reward and transitions of state–action pairs encountered in potentially good policies.

3.1 Oracle Minimax Program

Suppose we have already allocated some samples and have computed the maximum likelihood MDP $\widehat{\mathcal{M}}_k$ with empirical optimal policy $\widehat{\pi}_k^\star$ and know that the optimal value function estimate is at least ϵ_k-accurate, i.e., $\|V^\star - \widehat{V}_k^\star\|_\infty \le \epsilon_k$. How should we allocate further sampling resources to improve the accuracy in the optimal value function estimate? The idea is given by the simulation lemma (lemma 2): in order to see an improvement after sampling (i.e., in the next episode $k+1$) the maximum likelihood MDP $\widehat{\mathcal{M}}_{k+1}$ must have smaller confidence intervals $CI_{sa}(n_{sa}^{k+1})$ in the (s,a) pairs visited by $\pi^\star$ and the empirical optimal policy $\widehat{\pi}_{k+1}^\star$ on $\widehat{\mathcal{M}}_{k+1}$. Both are of course unknown. However, we introduce the constraint $(\widehat{V}_k^\star - \widehat{V}_k^\pi)(\rho) \le C\epsilon_k$ that restricts sampling to $C\epsilon_k$-optimal policies (and starting distributions) on $\widehat{\mathcal{M}}_k$. Here, C is a numerical constant that *will ensure that $\pi^\star$ and $\widehat{\pi}_{k+1}^\star$ satisfy this condition and are therefore allocated enough samples*. Given C and ϵ_k, the idea is that we should choose a sampling strategy $\{n_{sa}\}_{sa}$ with high enough samples to ensure $\sum_{(s,a)} \widehat{\widehat{w}}^{\pi,k+1,\rho} CI_{sa}(n_{sa}^{k+1}) \le \epsilon_{k+1}$ for all policies that satisfy $(\widehat{V}_k^\star - \widehat{V}_k^\pi)(\rho) \le C\epsilon_k$ so that Lemma 2 ensures $|(V^\star - \widehat{V}_{k+1}^\star)(\rho)| \le \epsilon_{k+1} = \epsilon_k/2$. This is equivalent to solving the following[1]:

Definition 1 (Oracle Minimax Problem)**.**

$$\min_n \max_{\pi,\rho} \quad \sum_{(s,a)} \widehat{\widehat{w}}_{sa}^{\pi,k+1,\rho} CI_{sa}(n_{sa}^{k+1}), \quad \textit{s.t.} \quad (\widehat{V}_k^\star - \widehat{V}_k^\pi)(\rho) \le C\epsilon_k, \sum_{(s,a)} n_{sa} \le n_{max}. \tag{1}$$

[1]For space, we omit the constraint $\rho_s \ge 0$ and $\|\rho\|_1 = 1$ on the starting distribution.

Here the vector of the discounted sum of visit probabilities $\widehat{\overline{w}}^{\pi,k+1,\rho}$ is computable from the linear system $(I-\gamma(\widehat{P}^{\pi}_{k+1})^{\top})\widehat{\overline{w}}^{\pi,k+1,\rho} = \rho$ and n_{max} is a guess on the number of samples needed to ensure that the objective function is $\leq \epsilon_k/2$. We call this problem the *oracle minimax problem* because it uses the next-episode empirical visit probabilities $\widehat{\overline{w}}^{\pi,k+1,\rho}$ which are not known. In addition, it uses the true variance of the next state value function (embedded in the definition of confidence intervals $CI_{sa}(n^k_{sa})$). As these quantities are unknown in episode k, the program cannot be solved.

3.2 Algorithm Minimax Program

This section shows how to construct a minimax program that is 'close' enough to the Oracle minimax problem (Equation 1) but is function of only empirical quantities computable from $\widehat{\mathcal{M}}_k$. The idea is 1) to avoid using the next-episode empirical distribution $\widehat{\overline{w}}^{\pi,k+1,\rho}$ and instead use the currently-computable $\widehat{\overline{w}}^{\pi,k,\rho}$ and 2) use the empirical variance of the next state value function $\mathrm{Var}_{\widehat{p}_k(s,a)} \widehat{V}^{\star}_k$ instead of the real, unknown variance $\mathrm{Var}_{p(s,a)} V^{\star}$. Estimating the visit distribution $\widehat{\overline{w}}^{\pi,k+1,\rho}$ accurately leads to a high sample complexity; fortunately we are able to claim that the product between *the visit distribution shift* $\widehat{\overline{w}}^{\pi,k+1,\rho} - \widehat{\overline{w}}^{\pi,k,\rho}$ *and the confidence interval vector* CI^{k+1} *on the rewards and transitions* is already small if policy π has received enough samples along its trajectories before the current episode. Let us rewrite the objective function of equation 1 as a function of the visit distribution on $\widehat{\mathcal{M}}_k$ plus a term that takes into account the shift in the distribution from $\widehat{\mathcal{M}}_k$ to $\widehat{\mathcal{M}}_{k+1}$:

$$\sum_{(s,a)} \widehat{\overline{w}}^{\pi,k+1,\rho}_{sa} CI_{sa}(n^{k+1}_{sa}) = \sum_{(s,a)} \underbrace{\widehat{\overline{w}}^{\pi,k,\rho}_{sa}}_{\text{Computable}} CI_{sa}(n^{k+1}_{sa}) + \sum_{(s,a)} \underbrace{(\widehat{\overline{w}}^{\pi,k+1,\rho}_{sa} - \widehat{\overline{w}}^{\pi,k,\rho}_{sa})}_{\text{Shift of Empirical Distributions}} CI_{sa}(n^{k+1}_{sa})$$

Lemma 9 in appendix allows us to claim that the rightmost summation above is less than $2Cp(n_{min})\epsilon_k$ for both[2] $\pi = \pi^{\star}$ and $\widehat{\pi}^{\star}_{k+1}$. Here $Cp(n_{min})$ is defined in appendix A and can be made (see lemma 16) for example $< 1/100$ by allocating a small constant number of samples $\tilde{O}(S/(1-\gamma)^2)$ to each (s,a) pair[3], independent of the desired accuracy ϵ_{k+1}. This way we can ensure that we can use $\widehat{\overline{w}}^{\pi,k,\rho}$ instead of $\widehat{\overline{w}}^{\pi,k+1,\rho}$ plus a small correction term $\ll \epsilon_k$.

Now the only quantities that are not known by the algorithm are the variance of the transitions and rewards that appear in the confidence intervals $CI_{sa}(n^{k+1}_{sa})$. Precisely, to estimate the variance of the transitions $\mathrm{Var}_{p(s,a)} V^{\star}$ in the (s,a) pair, we need to known both the transition probability $p(s,a)$ and the true value function $V^{\star}$, both of which are unknown. Fortunately it is possible to use the empirical transitions $\widehat{p}_k(s,a)$ and the empirical value function $\widehat{V}^{\star}_k$ plus a fast-shrinking (as a function of the number of samples) correction term. Since this analysis was similarly performed in prior papers for this setting [GMK13; Sid+18], we defer its discussion to Lemma 11 in the appendix. With these corrections (B_{ksa}, defined in appendix A, is the variance correction and $2\epsilon_k/625$ accounts for the distribution shift) we can write the following minimax problem:

Definition 2 (Algorithm Minimax Problem)**.**

$$\min_{n} \max_{\pi,\rho} \quad \sum_{(s,a)} \widehat{\overline{w}}^{\pi,k,\rho}_{sa} (\widehat{CI}_{sa}(n^{k+1}_{sa}) + B_{ksa}) + 2\epsilon_k/625$$
$$\textit{s.t.:} \quad (\widehat{V}^{\star}_k - \widehat{V}^{\pi}_k)(\rho) \leq C\epsilon_k; \quad \sum_{(s,a)} n_{sa} \leq n_{max}; \quad (I-\gamma(\widehat{P}^{\pi}_k)^{\top})\widehat{\overline{w}}^{\pi,k,\rho} = \rho. \qquad (2)$$

Here $\widehat{CI}_{sa}(n^{k+1}_{sa})$ are the confidence intervals evaluated with the empirical variances defined in Appendix A. This program is fully expressed in terms of empirical quantities that depends on $\widehat{\mathcal{M}}_k$. As long as a solution to the above minimax program is $\leq \epsilon_k/2$ the oracle objective function will also

[2] Lemma 9 bounds this term as $2Cp(n_{min})\epsilon^{\pi}_k$ for $\pi = \pi^{\star}, \pi = \widehat{\pi}^{\star}_{k+1}$, respectively; ϵ^{π}_k is defined in appendix A and represents the "accuracy" of policy π in episode k. To ensure $\epsilon^{\pi}_k \leq \epsilon_k$ we need an inductive argument which is sketched out in the main theorem (Theorem 1).

[3] As we will shortly see, this will contribute only a constant term to the final sample complexity.

be $\leq \epsilon_k/2$ at the solution of the program (for more details see Lemma 6 in the Appendix). In other words, by solving the minimax program (def 2) we put enough samples to satisfy the oracle program 1, which ensures accuracy in the value function estimate through Lemma 2.

4 Algorithm

Algorithm 1 BESPOKE

Input: Failure probability $\delta > 0$, accuracy $\epsilon_{Input} > 0$
Set $\epsilon_1 = \frac{1}{1-\gamma}$ and allocate n_{min} samples to each (s, a) pair
for $k = 1, 2, \ldots$
  for $n_{max} = 2^0, 2^1, 2^2, \ldots$
    Solve the optimization program of definition 7 (appendix)
    if the optimal value of the program of definition 7 is $\leq \frac{\epsilon_k}{2}$
      Break and return sampling strategy $\{n_{sa}^{k+1}\}_{sa}$
  Query the generative model up to $n_{sa}^{k+1}, \forall(s, a)$
  Compute, $\widehat{\pi}^\star_{k+1}$ and $\widehat{V}^\star_{k+1}$
  Set $\epsilon_{k+1} = \frac{\epsilon_k}{2}$
  if $\epsilon_{k+1} \leq \epsilon_{Input}$
    Break and return the policy $\widehat{\pi}^\star_{k+1}$

We now take the sampling approach described in the previous section and use it to construct an iterative algorithm for quickly learning a near-optimal or optimal policy given access to a generative model. Specifically we present *BESt POlicy identification with no Knowledge of the Environment* (BESPOKE) in Algorithm 1. The algorithm proceeds in episodes. Each episode starts with an empirical MDP $\widehat{\mathcal{M}}_k$ whose optimal value function $\widehat{V}^\star_k$ is ϵ_k accurate $\|V^\star - \widehat{V}^\star_k\|_\infty \leq \epsilon_k$ under an inductive assumption. The samples are allocated at each episode k by solving an optimization program equivalent to that in definition 2 to halve the accuracy in the value function estimate, i.e., $\|V^\star - \widehat{V}^\star_{k+1}\|_\infty \leq \epsilon_{k+1} = \epsilon_k/2$. In the innermost loop of the algorithm the required number of samples for the next episode is guessed $n_{max} = 1, 2, 4, 8, \ldots$, until n_{max} is sufficient to ensure that the objective function of the minimax problem of definition 2 will be $\leq \epsilon_k/2$; the purpose of the inner loop is to avoid putting more samples than needed and allows us to obtain the sample complexity result of Theorem 2. In Appendix G we reformulate the optimization program 2 (described more precisely in Definition 5 in the appendix) obtaining a convex minimization program that avoids optimizing over the policy and instead works directly with the distribution $\widehat{\overline{w}}^{\pi,k,\rho}$; this can be efficiently solved with standard techniques from convex optimization [BV04].

Theorem 1 (BESPOKE Works as Intended)**.** *With probability at least $1 - \delta$, in every episode k* BESPOKE *maintains an empirical MDP $\widehat{\mathcal{M}}_k$ such that its optimal value function $\widehat{V}^\star_k$ and its optimal policy $\widehat{\pi}^\star_k$ satisfy:*

$$\|V^\star - \widehat{V}^\star_k\|_\infty \leq \epsilon_k, \quad \|V^\star - V^{\widehat{\pi}^\star_k}\|_\infty \leq 2\epsilon_k$$

where $\epsilon_{k+1} \overset{def}{=} \frac{\epsilon_k}{2}$, $\forall k$. In particular, when BESPOKE *terminates in episode k_{Final} it holds that $\frac{\epsilon_{Input}}{2} \leq \epsilon_{k_{Final}} \leq \epsilon_{Input}$.*

The proof is reported in the appendix, and shows by induction that for every episode k, $\pi^\star$ and $\widehat{\pi}^\star_{k+1}$ are in the set of 'candidate' policies because they are near-optimal on $\widehat{\mathcal{M}}_k$, satisfying $(\widehat{V}^\star_k - \widehat{V}^{\pi^\star}_k)(\rho) \leq C\epsilon_k$ and $(\widehat{V}^\star_k - \widehat{V}^{\widehat{\pi}^\star_{k+1}}_k)(\rho) \leq C\epsilon_k$ for all ρ and are therefore allocated enough samples; this guarantees accuracy in $\widehat{V}^\star_{k+1}$ by Lemma 2.

5 Sample Complexity Analysis

To analyze the sample complexity of BESPOKE we derive another optimization program function of only problem dependent quantities. We 1) shift from the empirical visit distribution $\widehat{\overline{w}}^{\pi,k,\rho}$ on $\widehat{\mathcal{M}}_k$ to the "real" visit distribution $\overline{w}^{\pi,\rho}$ on $\mathcal{M}$; 2) move from empirical confidence intervals to those evaluated with the true variances; and finally 3) relax the near-optimality constraint on the policies by using Lemma 7 in the appendix (for an appropriate numerical constant $C^\star > C$) in order to be able to use the value functions on $\mathcal{M}$:

$$(\widehat{V}^\star_k - \widehat{V}^\pi_k)(\rho) \leq C\epsilon_k \rightarrow (V^\star - V^\pi)(\rho) \leq C^\star\epsilon_k, \quad \forall\rho. \tag{3}$$

In the end, we have enlarged the feasible set of the algorithm minimax problem while upper bounding its objective function obtaining:[4]

Definition 3 ($\star$-Minimax Program)**.**

$$\min_{n} \max_{\overline{w}^{\pi,\rho}} \sum_{(s,a)} \overline{w}^{\pi,\rho}_{sa}(CI_{sa}(n^{k+1}_{sa}) + 2B_{ksa}) + \epsilon_k/25 \tag{4}$$

subject to the constraints ($r \in \mathbb{R}^{SA}$ is the reward vector):

$$\underbrace{(V^\star - V^\pi)(\rho)}_{V^\star(\rho)-(\overline{w}^{\pi,\rho})^\top r} \leq C^\star \epsilon_k; \quad \sum_{(s,a)} n_{sa} \leq n_{max}; \quad (I - \gamma(P^\pi)^\top)\overline{w}^{\pi,\rho} = \rho. \tag{5}$$

This is made rigorous in Lemma 6, but essentially we have obtained a minimax program whose solution can be studied in terms of problem dependent quantities; in particular, its solution in terms of number of samples n_{sa} upper bounds the sample complexity of the algorithm in every episode.

Problem Dependent Analysis Due to space constraints, here we sketch the sample complexity analysis of suboptimal actions to make the gaps $\Delta_{sa} \overset{def}{=} V^\star(s) - Q^\star(s,a)$ appear while simultaneously eliminating the horizon dependence. We recall the following (e.g., Lemma 5.2.1 in [Kak+03]; see also our appendix):

Lemma 1 (Sum of Losses)**.** *It holds that:*

$$(V^\star - V^\pi)(\rho) = \sum_{(s,a)} \overline{w}^{\pi,\rho}_{sa}(\underbrace{Q^\star(s,\pi^\star(s)) - Q^\star(s,a)}_{\overset{def}{=}\Delta_{sa}}) = \sum_{(s,a)} \overline{w}^{\pi,\rho}_{sa}\Delta_{sa} \tag{20}$$

Lemma 1 expresses the value of a suboptimal policy as a sum of per-step losses Δ_{sa} weighted by the discounted sum of probabilities of being in that (s,a) pair. The key step that enables us to obtain strong problem dependent bounds and to remove the horizon dependence for suboptimal actions is synthesized in the following short lemma, where we ignore the term $(\sum_{(s,a)} 2\overline{w}^{\pi,\rho}_{sa}B_{ksa} + 3\epsilon_k/625)$.

Lemma 1 (Gap-Confidence Interval Lemma)**.** *If (π,ρ) satisfies $(V^\star - V^\pi)(\rho) \leq C^\star\epsilon_k$ then a sample complexity:*

$$n_{sa} = \tilde{O}\left(\underbrace{\frac{\operatorname{Var} R(s,a)}{\Delta^2_{sa}} + \frac{1}{\Delta_{sa}}}_{\text{Reward Estimation}} + \underbrace{\frac{\gamma^2 \operatorname{Var}_{p(s,a)} V^\star}{\Delta^2_{sa}} + \frac{\gamma}{(1-\gamma)\Delta_{sa}}}_{\text{Transition Estimation}}\right), \quad \forall (s,a) \tag{6}$$

suffices to ensure

$$\max_{\overline{w}^{\pi,\rho}} \sum_{(s,a)} \overline{w}^{\pi,\rho}_{sa} CI_{sa}(n^{k+1}_{sa}) \leq \frac{\epsilon_k}{2}. \tag{7}$$

Proof. A direct computation shows that if n^{k+1}_{sa} satisfies equation 6 with appropriate constants[5] then:

$$CI_{sa}(n^{k+1}_{sa}) \leq \frac{\Delta_{sa}}{2C^\star}. \tag{8}$$

This justifies the first inequality below:

$$\sum_{(s,a)} \overline{w}^{\pi,\rho}_{sa} CI_{sa}(n^{k+1}_{sa}) \leq \frac{1}{2C^\star}\sum_{(s,a)} \overline{w}^{\pi,\rho}_{sa}\Delta_{sa} = \frac{1}{2C^\star}(V^\star - V^\pi)(\rho) \leq \frac{1}{2}\epsilon_k. \tag{9}$$

The equality follows from lemma 1 and the last inequality from the constraint on the optimality of π. □

[4]The relaxed optimization program is over the distribution induced by the policy. Here, P^π is the transition matrix identified by the policy π on $\mathcal{M}$.

[5]Note that, in particular, $C^\star$ is a constant.

They key idea is that by *having confidence intervals of the same size as the gaps is sufficient to estimate the policy as accurately as its suboptimality gap* $(V^\star - V^\pi)(\rho)$, regardless of the horizon. By augmenting this argument with the law of total variance [GMK13], splitting into further subcases, and by taking into account the correction terms we obtain:

Theorem 2 (Sample Complexity of the Algorithm BESPOKE)**.** *With probability at least* $1-\delta$*, the total sample complexity of* BESPOKE *up to episode* k *is upper bounded by* $\sum_{(s,a)} n_{sa}$ *where* n_{sa} *is the total number of samples allocated to the* (s,a) *pair:*

$$n_{sa} = \tilde{O}\Big(\min\Big\{ \frac{1}{(1-\gamma)^3(\epsilon_k)^2}, \frac{\operatorname{Var} R(s,a) + \gamma^2 \operatorname{Var}_{p(s,a)} V^\star}{(1-\gamma)^2(\epsilon_k)^2} + \frac{1}{(1-\gamma)^2(\epsilon_k)}, \tag{166}$$

$$\frac{\operatorname{Var} R(s,a) + \gamma^2 \operatorname{Var}_{p(s,a)} V^\star}{\Delta_{s,a}^2} + \frac{1}{(1-\gamma)\Delta_{s,a}} \Big\} + \frac{\gamma S}{(1-\gamma)^2} \Big). \tag{167}$$

Notice that the BESPOKE would suffer a worst-case sample complexity similar to [GMK13; Sid+18] only in the initial phases of learning, i.e., whenever ϵ_k is much larger than the gaps.

6 Significance of the Bound

We motivate the importance of theorem 2 by specializing the result in two noteworthy cases.

Sample Complexity to Identify the Best Policy and the Worst-Case Lower Bound If the optimal policy is unique, define the minimum gap $\Delta_{min} = \min_{s,a,a\neq\pi^\star(a)} \Delta_{sa}$. To identify the optimal policy we must set $\epsilon_{Input} \leq \Delta_{min}$ and the sample complexity of BESPOKE at termination becomes:

$$\tilde{O}\Bigg(\underbrace{\sum_s \min\left\{ \frac{1}{(1-\gamma)^3\Delta_{min}^2}, \frac{\operatorname{Var} R(s,\pi^\star(s)) + \gamma^2 \operatorname{Var}_{p(s,\pi^\star(s))} V^\star}{(1-\gamma)^2\Delta_{min}^2} + \frac{1}{(1-\gamma)^2\Delta_{min}} \right\}}_{\text{ESTIMATING } \pi^\star}$$

$$+ \underbrace{\sum_{(s,a)\mid a\neq\pi^\star(s)} \left(\frac{\operatorname{Var} R(s,a) + \gamma^2 \operatorname{Var}_{p(s,a)} V^\star}{\Delta_{sa}^2} + \frac{1}{(1-\gamma)\Delta_{sa}} \right)}_{\text{RULING-OUT SUBOPTIMAL ACTIONS}} + \underbrace{\frac{\gamma S^2 A}{(1-\gamma)^2}}_{\text{CONSTANT}} \Bigg) \tag{10}$$

One of our core contributions is that we suffer a dependence on the horizon $1/(1-\gamma)$ only in estimating the optimal (s,a) pairs (first summation over the state space). *The summation over suboptimal* (s,a) *is independent of the horizon*, although of the horizon implicitly affects the scaling of the variance $\operatorname{Var}_{p(s,a)} V^\star$ and explicitly the maximum value function range (term $1/(1-\gamma)\Delta_{sa}$).

It is important to compare the above result with the established lower bound [GMK13] which is $\Omega(\frac{N}{(1-\gamma)^3\epsilon^2})$ to obtain an ϵ-accurate policy, where N is the number of state-action pairs. Since $\Delta_{sa} = \Delta_{min}, \ \forall (s,a), a \neq \pi^\star(s)$ in the lower bound construction and the variance is maximum $\operatorname{Var}_{p(s,a)} V^\star \leq 1/(1-\gamma)^2$, we are able to identify the optimal policy in $\tilde{O}\left(\frac{S}{(1-\gamma)^3\Delta_{min}^2} + \frac{S(A-1)}{(1-\gamma)^2\Delta_{min}^2} + \frac{S^2A}{(1-\gamma)^2} \right)$ samples which improves[6] on the worst case bound $\tilde{O}\left(\frac{SA}{(1-\gamma)^3\Delta_{min}^2} + \frac{S^2A}{(1-\gamma)^2} \right)$ of [GMK13; Sid+18] by a full horizon factor for suboptimal actions. While our result can be surprising at first, it does not contradict the lower bound: the lower bound makes no attempt to distinguish between optimal and suboptimal actions as it is only expressed in terms of *total* (s,a) *pairs* N, and the construction uses a number of (s,a) pairs that is a *constant* multiple of the state space cardinality, i.e., one could only deduce $\Omega(\frac{S}{(1-\gamma)^3\Delta_{min}^2})$ as a lower bound. Our result, therefore, *does not violate the lower bound*, but rather it shows that while we must suffer an unavoidable worst-case $1/(1-\gamma)^3$ factor on the state space corresponding to the optimal (s,a) pairs, the dependence on the planning horizon is absent for suboptimal (s,a) except for the scaling of the value function implicit in the variance. Surprisingly, excluding the constant term $\frac{S^2A}{(1-\gamma)^2}$, suboptimal (s,a) pairs get a combined number of samples

[6]The paper [Sid+18] has the same bound as [GMK13] but avoids the constant term $\frac{S^2A}{(1-\gamma)^2}$.

$\tilde{O}\left(\sum_{(s,a)|a\neq\pi^\star(s)}\left(\frac{\operatorname{Var} R(s,a)+\gamma^2 \operatorname{Var}_{p(s,a)} V^\star}{\Delta_{sa}^2}+\frac{1}{(1-\gamma)\Delta_{sa}}\right)\right)$ which is the sample complexity (ignoring log and constant factors) that a variance-aware bandit algorithm for best arm identification would need (see e.g., [GGL12], appendix B) to 'reject' these suboptimal arms provided that it can obtain samples[7] of the random variable $R(s,a)+\gamma V^\star(s'),\ s'\sim p(s,a)$. In this case, however, the $V^\star$ vector would need to be known to the bandit algorithm. In other words, the sample complexity of BESPOKE at termination consists of two main terms: a leading order term with a dependence on the state space with an unavoidable (due to the lower bound) dependence on the horizon $\frac{1}{1-\gamma}$, and an horizon-free bandit-like sample complexity to rule out suboptimal actions as if the optimal value function $V^\star$ was known.

BESPOKE applied to Bandits Finally, if $\gamma = 0$ we are in the bandit setting, and the sample complexity of BESPOKE at step k becomes exactly (since $\operatorname{Var} R(s,a) \le 1$):

$$\tilde{O}\left(\sum_{(s,a)}\left(\frac{\operatorname{Var} R(s,a)}{\max\{\epsilon_k^2,\Delta_{sa}^2\}}+\frac{1}{\max\{\epsilon_k,\Delta_{sa}\}}\right)\right)\le \tilde{O}\left(\sum_{(s,a)}\frac{1}{\max\{\epsilon_k^2,\Delta_{sa}^2\}}\right) \tag{11}$$

This matches the best-known sample complexity bound for best arm identification for tabular bandit with gaps and variances [ABM10; GGL12] except for constants and log terms. This is encouraging as it suggests it may be possible to have algorithms with a smooth transition in sample complexity as a function of the discount factor when moving from a bandit to an RL setting.

7 Related Literature and Conclusion

Related Literature In the more challenging setting of online exploration (i.e., without a generative model) the PAC literature [DB15; DLB17; LH14; SLL09] directly provides algorithms to identify an ϵ-optimal policy with high probability in the worst-case. Gap-aware analyses exists, see for example [BK97; TB08; OPT18] for asymptotic regret bounds on ergodic MDPs with matching upper and lower bounds and with an emphasis on the minimum gap; since these analyses look at the asymptotic regret they are not comparable to the proposal here. Very recently [SJ19] presents a gap-based non-asymptotic regret bound for episodic MDPs but not yet free of the horizon and dependencies on Δ_{min}. Gaps in MDPS have also been used to justify the observed relation between the value function accuracy and the resulting policy performance [FSM10]. In addition, [EMM06; Bru10] also propose an algorithm and PAC bounds that depend the minimum gap, but the results do not leverage recent advances in tighter sample complexity analysis. [JOA10] presents a regret bound based on the same quantity. The maximum variance of the next-state optimal value function is discussed in [MMM14; ZB19].

The closest related work in the PAC setting similarly assumes access to a generative model, and provides near-matching worst-case sample complexity upper and lower bounds [AMK12] for tabular MDPs even in terms of computational complexity [Sid+18]. However, this work focuses on near-optimal worst-case performance: as these algorithms allocate samples uniformly they do not adapt to the problem structure. Finally, [Aga+19] show how to improve on the constant sample complexity term for model based approaches like the one we use here; it is possible that their techniques can be applied to our setting.

Conclusion This work leverages domain structure, notably the action-value function gaps, to eliminate the impact of the horizon when ruling out suboptimal actions to identify a near-optimal policy for discounted-reward Markov decision processes using a generative model, except for a constant term and the inherent value function scaling. This is achieved through a tractable algorithm. In doing so, our finite time sample complexity analysis quantifies the sample complexity contribution of each state-action pair as a function of the action-value function gaps and variances of the rewards and next-state value function, and recovers the best-known bounds (excepts for logs and constants) when deployed to bandit instances using these quantities.

Our work provides at least two important analytical tools: 1) the way we relate the suboptimality of the policies with the gaps to reduce the dependence on the horizon is new, and could be used in

[7] Here, $\operatorname{Var} R(s,a)+\gamma^2\operatorname{Var}_{p(s,a)} V^\star$ is the variance of the random variable $R(s,a)+\gamma V^\star(s')$ with $s'\sim p(s,a)$. Note the scaling of this random variable, which has range $\frac{1}{1-\gamma}$.

other settings to make the gap appear while simultaneously reducing the horizon dependence 2) the way we analyze the visit distribution shift induced by the policies, weighted by the local reward and transition confidence intervals, and show it is small, is another analytical contribution of our work which can be extended to the settings where one is interested in obtaining a good policy from a given starting distribution ρ as opposed to all starting states.

Acknowledgment

This work is partially supported by a Total Innovation Fellowship program, an NSF CAREER award and an Office of Naval Research Young Investigator Award.

References

[ABF02] Peter Auer, Nicolo Cesa Bianchi, and Paul Fischer. "Finite-time Analysis of the Multiarmed Bandit Problem". In: *Machine Learning* (2002).

[ABM10] Jean-Yves Audibert, Sebastien Bubeck, and Remi Munos. "Best Arm Identification in Multi-Armed Bandits". In: *Conference on Learning Theory (COLT)*. 2010.

[Aga+19] Alekh Agarwal et al. "Optimality and approximation with policy gradient methods in markov decision processes". In: *arXiv preprint arXiv:1908.00261* (2019).

[AMK12] Mohammad Gheshlaghi Azar, Remi Munos, and Hilbert J. Kappen. "On the Sample Complexity of Reinforcement Learning with a Generative Model". In: *International Conference on Machine Learning (ICML)*. 2012.

[AMS09] Jean Yves Audibert, Remi Munos, and Csaba Szepesvari. "Exploration-exploitation trade-off using variance estimates in multi-armed bandits". In: *Theoretical Computer Science* (2009).

[AOM17] Mohammad Gheshlaghi Azar, Ian Osband, and Remi Munos. "Minimax Regret Bounds for Reinforcement Learning". In: *International Conference on Machine Learning (ICML)*. 2017.

[BK97] Apostolos N Burnetas and Michael N Katehakis. "Optimal adaptive policies for Markov decision processes". In: *Mathematics of Operations Research* 22.1 (1997), pp. 222–255.

[BMS09] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. "Pure exploration in multi-armed bandits problems". In: *International Conference on Algorithmic Learning Theory*. 2009.

[Bru10] Emma Brunskill. "When Policies Can Be Trusted: Analyzing a Criteria to Identify Optimal Policies in MDPs with Unknown Model Parameters." In: *International Conference on Automated Planning and Scheduling (ICAPS)*. 2010, pp. 218–221.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[DB15] Christoph Dann and Emma Brunskill. "Sample Complexity of Episodic Fixed-Horizon Reinforcement Learning". In: *Advances in Neural Information Processing Systems (NIPS)*. 2015.

[DLB17] Christoph Dann, Tor Lattimore, and Emma Brunskill. "Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning". In: *Advances in Neural Information Processing Systems (NIPS)*. 2017.

[EMM06] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. "Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems". In: *Journal of Machine Learning Research* (2006).

[FSM10] Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. "Error propagation for approximate policy and value iteration". In: *Advances in Neural Information Processing Systems (NIPS)*. 2010.

[GGL12] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. "Best arm identification: A unified approach to fixed budget and fixed confidence". In: *Advances in Neural Information Processing Systems (NIPS)*. 2012, pp. 3212–3220.

[GMK13] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model". In: *Machine Learning* 91.3 (June 2013), pp. 325–349. DOI: 10.1007/s10994-013-5368-1. URL: https://doi.org/10.1007/s10994-013-5368-1.

[Jam+14] Kevin Jamieson et al. "lil'ucb: An optimal exploration algorithm for multi-armed bandits". In: *Conference on Learning Theory (COLT)*. 2014, pp. 423–439.

[JOA10] Thomas Jaksch, Ronald Ortner, and Peter Auer. "Near-optimal Regret Bounds for Reinforcement Learning". In: *Journal of Machine Learning Research* (2010).

[Kak+03] Sham Machandranath Kakade et al. "On the sample complexity of reinforcement learning". PhD thesis. University of London London, England, 2003.

[KKS13] Zohar Karnin, Tomer Koren, and Oren Somekh. "Almost optimal exploration in multi-armed bandits". In: *International Conference on Machine Learning (ICML)*. 2013.

[KMN02] Michael Kearns, Yishay Mansour, and Andrew Y Ng. "A sparse sampling algorithm for near-optimal planning in large Markov decision processes". In: *Machine Learning* 49.2-3 (2002), pp. 193–208.

[LH14] Tor Lattimore and Marcus Hutter. "Near-optimal PAC bounds for discounted MDPs". In: *Theoretical Computer Science* 558 (2014), pp. 125–143.

[MM94] Oded Maron and Andrew W Moore. "Hoeffding races: Accelerating model selection search for classification and function approximation". In: *Advances in Neural Information Processing Systems (NIPS)*. 1994, pp. 59–66.

[MMM14] Odalric-Ambrym Maillard, Timothy A. Mann, and Shie Mannor. ""How hard is my MDP?" The distribution-norm to the rescue". In: *Advances in Neural Information Processing Systems (NIPS)*. 2014.

[MP09] Andreas Maurer and Massimiliano Pontil. "Empirical Bernstein Bounds and Sample Variance Penalization". In: *Conference on Learning Theory (COLT)*. 2009.

[MSA08] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. "Empirical bernstein stopping". In: *International Conference on Machine Learning (ICML)*. ACM. 2008, pp. 672–679.

[OPT18] Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. "Exploration in Structured Reinforcement Learning". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8874–8882.

[OVR13] Ian Osband, Benjamin Van Roy, and Daniel Russo. "(More) Efficient Reinforcement Learning via Posterior Sampling". In: *Advances in Neural Information Processing Systems (NIPS)*. 2013.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, 2018.

[Sid+18] Aaron Sidford et al. "Near-Optimal Time and Sample Complexities for for Solving Discounted Markov Decision Process with a Generative Model". In: *Advances in Neural Information Processing Systems (NIPS)*. 2018.

[SJ19] Max Simchowitz and Kevin Jamieson. "Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs". In: *arXiv preprint arXiv:1905.03814* (2019).

[SLL09] Alexander L Strehl, Lihong Li, and Michael L Littman. "Reinforcement learning in finite MDPs: PAC analysis". In: *Journal of Machine Learning Research* 10.Nov (2009), pp. 2413–2444.

[TB08] Ambuj Tewari and Peter L Bartlett. "Optimistic linear programming gives logarithmic regret for irreducible MDPs". In: *Advances in Neural Information Processing Systems*. 2008, pp. 1505–1512.

[WBS07] Tao Wang, Michael Bowling, and Dale Schuurmans. "Dual representations for dynamic programming and reinforcement learning". In: *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*. IEEE. 2007, pp. 44–51.

[Wei+03] Tsachy Weissman et al. *Inequalities for the l1 deviation of the empirical distribution*. Tech. rep. Hewlett-Packard Labs, 2003.

[ZB19] Andrea Zanette and Emma Brunskill. "Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds". In: *International Conference on Machine Learning (ICML)*. 2019. URL: `http://proceedings.mlr.press/v97/zanette19a.html`.

A Symbols

Table 1: Notation

δ'_n	$\overset{def}{=}$	$\frac{\delta}{4SAn^2}$
c_n	$\overset{def}{=}$	$\ln(4/\delta'_n)$
ρ	$\overset{def}{=}$	starting distribution
$\mathcal{M}$	$\overset{def}{=}$	true MDP
$\widehat{\mathcal{M}}_k$	$\overset{def}{=}$	empirical MDP at step k
$V^\star$	$\overset{def}{=}$	optimal value function on $\mathcal{M}$
$\widehat{V}^\star_k$	$\overset{def}{=}$	optimal value function on $\widehat{\mathcal{M}}_k$
$\pi^\star$	$\overset{def}{=}$	optimal policy on $\mathcal{M}$
$\widehat{\pi}^\star_k$	$\overset{def}{=}$	empirical optimal policy on $\widehat{\mathcal{M}}_k$
$\overline{w}^{\pi,\rho}_{sa}$	$\overset{def}{=}$	visit probability to (s,a) on $\mathcal{M}$ upon following π with starting distribution ρ
$\widehat{\overline{w}}^{\pi,k,\rho}_{sa}$	$\overset{def}{=}$	visit probability to (s,a) on $\widehat{\mathcal{M}}_k$ upon following π with starting distribution ρ
$V(\rho)$	$\overset{def}{=}$	$\sum_i \rho_i V(s_i)$
Δ_{sa}	$\overset{def}{=}$	$Q^\star(s,\pi^\star(s)) - Q^\star(s,a)$
$BI(\sigma,b,n)$	$\overset{def}{=}$	$\sqrt{\frac{2c_n\sigma^2}{n}} + \frac{bc_n}{3(n-1)}$ (Bernstein Inequality)
$CI_{sa}(n^k_{sa})$	$\overset{def}{=}$	$BI(\mathrm{Var}\, R(s,a), 1, n^k_{sa}) + \gamma BI(\mathrm{Var}_{p(s,a)} V^\star, \frac{1}{(1-\gamma)}, n^k_{sa})$
$\widehat{CI}_{sa}(n^k_{sa})$	$\overset{def}{=}$	$BI(\mathrm{Var}\, \widehat{R}(s,a), 1, n^k_{sa}) + \gamma BI(\mathrm{Var}_{\widehat{p}_k(s,a)} \widehat{V}^\star_k, \frac{1}{(1-\gamma)}, n^k_{sa})$
CI^k	$\overset{def}{=}$	Vector containing the $CI_{sa}(n^k_{sa})$ values
B_{ksa}	$\overset{def}{=}$	$\frac{2c_n}{(1-\gamma)(n_{sa}-1)} + \gamma\sqrt{\frac{2c_n}{n_{sa}}}\epsilon_k$
B_k	$\overset{def}{=}$	Vector containing the B_{ksa} values
ϵ^π_k	$\overset{def}{=}$	$\max_{\rho,\rho\geq 0,\Vert\rho\Vert_1=1} \sum_{(s,a)} \widehat{\overline{w}}^{\pi,k,\rho}_{sa} CI_{sa}(n^k_{sa})$
C	$\overset{def}{=}$	20
$C^\star$	$\overset{def}{=}$	$\frac{2+C}{1-Cp(n_{min})}$
$Cp(n_{min})$	$\overset{def}{=}$	$\frac{\gamma}{(1-\gamma)}\sqrt{\frac{2Sc_n}{n_{min}}}$
n	$\overset{def}{=}$	vector that contains the number of samples to each (s,a), unless it's a generic scalar
$\widehat{S}_k$	$\overset{def}{=}$	$\{\pi \mid (\widehat{V}^\star_k - \widehat{V}^\pi_k)(\rho) \leq C\epsilon_k, \quad \forall\rho\geq 0, \Vert\rho\Vert_1 = 1\}$
S_k	$\overset{def}{=}$	$\{\pi \mid (V^\star - V^\pi)(\rho) \leq C^\star\epsilon_k, \quad \forall\rho\geq 0, \Vert\rho\Vert_1 = 1\}$
n_{min}	$\overset{def}{=}$	$\frac{2\times 625^2\gamma^2 Sc_n}{(1-\gamma)^2}$

B Standard Concentration Inequalities

Proposition 1 (Bernstein's Inequality)**.** *Let $X_1, \cdots, X_n$ be i.i.d. random variables with values in $[0, C]$ and let $\delta' > 0$. Then with probability at least $1-\delta'$ in $(X_1, \ldots, X_n)$ we have:*

$$\left|\mathbb{E}\, X - \frac{1}{n}\sum_{i=1}^{n} X_i\right| \leq \sqrt{\frac{2\,\mathrm{Var}\, X \ln\left(2/\delta'\right)}{n}} + \frac{C\ln(2/\delta')}{3n}. \tag{12}$$

Proof. See [MP09], Theorem 3. □

Corollary 1 (Bernstein's inequality applied to Transition Probabilities)**.** *Let p be a k-dimensional transition probability vector (such that $\|p\|_1 = 1$) and let $\widehat{p}$ be its maximum likelihood estimate. Let $\delta' > 0$. Then with probability at least $1-\delta'$ we have:*

$$\left|\widehat{p}_i - p_i\right| \leq \sqrt{\frac{2p_i\ln\left(2k/\delta'\right)}{n}} + \frac{2\ln(2k/\delta')}{3n} \tag{13}$$

where p_i is the i-th component of p.

Proof. Immediate from Bernstein's inequality with the variance $p_i(1-p_i)$ of a Bernoulli random variable and a union bound over k. □

Proposition 2 (Converge Rate of Empirical Variance)**.** *Let $V^\star$ be a fixed vector with values in $[0, C]$ and let $\delta' > 0$:*

$$\left|\sqrt{\underset{\widehat{p}_k(s,a)}{\mathrm{Var}}\, V^\star} - \sqrt{\underset{p(s,a)}{\mathrm{Var}}\, V^\star}\right| \leq C\sqrt{\frac{2\ln\left(2/\delta'\right)}{n-1}}. \tag{14}$$

Proof. See [MP09], Theorem 10. □

Proposition 3 (Weissman et al. Inequality)**.** *Let $\widehat{p}$ be the maximum likelihood probability vector of the distribution p obtained by drawing i.i.d. samples from the discrete distribution p with k point masses, and let $\delta' > 0$. With probability at least $1-\delta'$ it holds that:*

$$\|\widehat{p} - p\|_1 \leq \sqrt{\frac{2k\log(2/\delta')}{n}} \overset{def}{=} \frac{\sqrt{2kc_n}}{\sqrt{n}}. \tag{15}$$

Proof. See [Wei+03]. □

Lemma 2 (Good Event)**.** *With probability at least $1-\delta$ the following events holds true for all (s, a) pairs for all episodes of the algorithm:*

$$|\widehat{r}_k(s,a) - r(s,a)| \leq \sqrt{\frac{2c_n\left(\mathrm{Var}\, R(s,a)\right)}{n_{sa}}} + \frac{c_n}{3(n_{sa}-1)} \tag{16}$$

$$|\left(\widehat{p}_k(s,a) - p(s,a)\right)^\top V^\star| \leq \sqrt{\frac{2c_n\left(\mathrm{Var}_{p(s,a)}\, V^\star\right)}{n_{sa}}} + \frac{c_n}{3(1-\gamma)(n_{sa}-1)} \tag{17}$$

$$\left|\sqrt{\underset{\widehat{p}_k(s,a)}{\mathrm{Var}}\, V^\star} - \sqrt{\underset{p(s,a)}{\mathrm{Var}}\, V^\star}\right| \leq \frac{1}{1-\gamma}\sqrt{\frac{2c_n}{n-1}} \tag{18}$$

$$\|\widehat{p}_k(s,a) - p(s,a)\|_1 \leq \frac{\sqrt{2Sc_n}}{\sqrt{n}} \tag{19}$$

Proof. Using propositions 1,2,3 and corollary 1 with a union bound over the (s, a) pairs and over the maximum number of samples n_{sa}. □

C Preliminaries

In this section we recall some standard results in reinforcement learning. The results have been adapted so that they could be expressed in terms of a starting distribution ρ instead of a fixed starting state.

C.1 Sum of Losses

The lemma below expresses the difference in values between two policies as a sum of the per-step losses:

Lemma 1 (Sum of Losses)**.** *It holds that:*

$$(V^\star - V^\pi)(\rho) = \sum_{(s,a)} \overline{w}_{sa}^{\pi,\rho} (\underbrace{Q^\star(s,\pi^\star(s)) - Q^\star(s,a)}_{\overset{def}{=}\Delta_{sa}}) = \sum_{(s,a)} \overline{w}_{sa}^{\pi,\rho} \Delta_{sa} \tag{20}$$

Proof. Consider a fixed starting state s. We have that:

$$\begin{aligned}(V^\star - V^\pi)(s) &= r(s,\pi^\star(s)) - r(s,\pi(s)) + \gamma p(s,\pi^\star(s))^\top V^\star - \gamma p(s,\pi(s))^\top V^\pi & (21)\\ &= r(s,\pi^\star(s)) - r(s,\pi(s)) + \gamma(p(s,\pi^\star(s)) - p(s,\pi(s)))^\top V^\star + & (22)\\ &+ \gamma p(s,\pi(s))^\top (V^\star - V^\pi) & (23)\end{aligned}$$

Induction with a ρ-weighted average over the starting state and the definition of $Q^\star$ values (and their gaps Δ_{sa}) conclude the proof. □

C.2 Simulation Lemmas

The Simulation Lemma below allows to evaluate policy π on two different MDPs, with the induced distribution evaluated on the empirical MDP and the value function for the backup on the real MDP.

Lemma 3 (Simulation Lemma)**.** *It holds that:*

$$(\widehat{V}_k^\pi - V^\pi)(\rho) = \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} \left(\widehat{r}_k(s,a) - r(s,a) + \gamma(\widehat{p}_k(s,a) - p(s,a))^\top V^\pi\right) \tag{24}$$

Proof. From any starting state s:

$$\begin{aligned}(\widehat{V}_k^\pi - V^\pi)(s) &= \widehat{r}_k(s,a) - r(s,a) + \gamma(\widehat{p}_k(s,a)^\top \widehat{V}_k^\pi - p(s,a)^\top V^\pi) & (25)\\ &= \widehat{r}_k(s,a) - r(s,a) + \gamma(\widehat{p}_k(s,a) - p(s,a))^\top V^\pi + \gamma \widehat{p}_k(s,a)^\top (\widehat{V}_k^\pi - V^\pi) & (26)\end{aligned}$$

Induction and a re-weighting by ρ concludes the proof.

□

The following lemma is a consequence of a lemma in [AMK12], and explains that to properly estimate the value function we need to estimate the rewards and transitions accurately only for the optimal policy and the optimal policy on the empirical MDP. Importantly, the lemma uses the true optimal value function $V^\star$.

Lemma 2 (Simulation Lemma for Optimal Value Function Estimate [GMK13])**.** *With probability at least $1-\delta$, outside the failure event for any starting distribution ρ it holds that:*

$$(V^\star - \widehat{V}_k^\star)(\rho) \le \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi^\star,k,\rho} \left((r - \widehat{r}_k)(s,a) + \gamma(p - \widehat{p}_k)(s,a)^\top V^\star\right) \le \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi^\star,k,\rho} CI_{sa}(n_{sa}^k)$$

$$(V^\star - \widehat{V}_k^\star)(\rho) \ge \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\widehat{\pi}_k^\star,k,\rho} \left((r - \widehat{r}_k)(s,a) + \gamma(p - \widehat{p}_k)(s,a)^\top V^\star\right) \ge -\sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\widehat{\pi}_k^\star,k,\rho} CI_{sa}(n_{sa}^k)$$

Proof. Lemma 2 in [AMK12] gives (here $\widehat{w}_{sa}^{\pi,k,s_0}$ is the discounted sum of visit probabilities upon starting from s_0 and following policy π on the empirical MDP $\widehat{\mathcal{M}}_k$):

$$V^\star(s_0) - \widehat{V}_k^\star(s_0) \le \sum_{(s,a)} \widehat{w}_{sa}^{\pi^\star,k,s_0} \left(\widehat{r}_k(s,a) - r(s,a) + \gamma\left(\widehat{p}_k(s,a) - p(s,a)\right)^\top V^\star\right) \tag{27}$$

$$V^\star(s_0) - \widehat{V}_k^\star(s_0) \ge \sum_{(s,a)} \widehat{w}_{sa}^{\widehat{\pi}_k^\star,k,s_0} \left(\widehat{r}_k(s,a) - r(s,a) + \gamma\left(\widehat{p}_k(s,a) - p(s,a)\right)^\top V^\star\right) \tag{28}$$

$$\tag{29}$$

Outside the failure event (lemma 2) it holds that:

$$V^\star(s_0) - \widehat{V}_k^\star(s_0) \le \sum_{(s,a)} \widehat{w}_{sa}^{\pi^\star,k,s_0} CI_{sa}(n_{sa}^k) \tag{30}$$

$$V^\star(s_0) - \widehat{V}_k^\star(s_0) \ge \sum_{(s,a)} \widehat{w}_{sa}^{\widehat{\pi}_k^\star,k,s_0} CI_{sa}(n_{sa}^k) \tag{31}$$

Finally, a weighted sum over the probabilities of starting at each starting state ρ_s yields the thesis. □

Next, we recall the following version of the simulation lemma that expresses the accuracy with which a generic policy π can be evaluated on the empirical vs real MDP as a function of its distance to the optimal value function.

Lemma 4 (Simulation Lemma for Policy Estimate)**.** *If n_{min} is the minimum number of samples allocated to any (s,a) pair then outside of the failure event it holds that:*

$$\|(\widehat{V}_k^\pi - V^\pi)\|_\infty \le \sum_{(s,a)} \widehat{w}_{sa}^{\pi,k,\rho} CI_{sa}(n_{sa}^k) + Cp(n_{min})\|\left(V^\pi - V^\star\right)\|_\infty. \tag{32}$$

Proof. Using the simulation lemma (lemma 3)

$$\begin{aligned}(\widehat{V}_k^\pi - V^\pi)(\rho) &= \sum_{(s,a)} \widehat{w}_{sa}^{\pi,k,\rho} \left(\widehat{r}_k(s,a) - r(s,a) + \gamma\left(\widehat{p}_k(s,a) - p(s,a)\right)^\top V^\pi\right) && (33)\\ &= \sum_{(s,a)} \widehat{w}_{sa}^{\pi,k,\rho} \left(\widehat{r}_k(s,a) - r(s,a) + \gamma\left(\widehat{p}_k(s,a) - p(s,a)\right)^\top V^\star\right) && (34)\\ &\quad + \gamma \sum_{(s,a)} \widehat{w}_{sa}^{\pi,k,\rho} \left(\widehat{p}_k(s,a) - p(s,a)\right)^\top \left(V^\pi - V^\star\right). && (35)\end{aligned}$$

Notice that we have the following upper bound outside of the failure event:

$$\Big| \sum_{(s,a)} \widehat{w}_{sa}^{\pi,k,\rho} \left(\widehat{r}_k(s,a) - r(s,a) + \gamma\left(\widehat{p}_k(s,a) - p(s,a)\right)^\top V^\star\right) \Big| \le \sum_{(s,a)} \widehat{w}_{sa}^{\pi,k,\rho} CI_{sa}(n_{sa}^k). \tag{36}$$

To bound the second line, if $n \ge n_{min}$ we obtain the upper bound (by Holder's inequality):

$$\le \frac{\gamma}{(1-\gamma)} \max_{(s,a)} \|\left(\widehat{p}_k(s,a) - p(s,a)\right)\|_1 \|\left(V^\pi - V^\star\right)\|_\infty \overset{def}{=} Cp(n_{min})\|\left(V^\pi - V^\star\right)\|_\infty. \tag{37}$$

□

C.3 Variance Lemma

The lemma below allows to express how much the variance varies when we consider different value functions.

Lemma 5. *For any two random variables V_1, V_2 it holds that:*

$$\left|\sqrt{\mathrm{Var}(V_1)} - \sqrt{\mathrm{Var}(V_2)}\right| \le \|V_1 - V_2\|_{2,p} \le \|V_1 - V_2\|_\infty \tag{38}$$

where $\|\cdot\|_{2,p}$ denotes the 2-norm of the random variables (i.e., the second moment) under p and $\|\cdot\|_\infty$ is the almost sure upper bound to the random variable.

Proof. Consider the mean-centered random variables $\overline{V}_1 = V_1 - \mathbb{E}\, V_1$ and $\overline{V}_2 = V_2 - \mathbb{E}\, V_2$. Then:

$$\sqrt{\mathrm{Var}(V_1)} = \sqrt{\mathrm{Var}(\overline{V}_1)} = \sqrt{\mathbb{E}(\overline{V}_1)^2} = \|\overline{V}_1\|_{2,p} = \|\overline{V}_2 + \overline{V}_1 - \overline{V}_2\|_{2,p} \tag{39}$$

$$\leq \|\overline{V}_2\|_{2,p} + \|\overline{V}_1 - \overline{V}_2\|_{2,p} = \sqrt{\mathbb{E}(\overline{V}_2)} + \sqrt{\mathbb{E}(\overline{V}_1 - \overline{V}_2)^2} \tag{40}$$

$$= \sqrt{\mathrm{Var}(\overline{V}_2)} + \sqrt{\mathbb{E}(V_1 - V_2)^2 - \left(\mathbb{E}(V_1 - V_2)\right)^2} \tag{41}$$

$$= \sqrt{\mathrm{Var}(V_2)} + \sqrt{\mathrm{Var}(V_1 - V_2)}. \tag{42}$$

where the inequality is Minkowski's inequality (i.e., the triangle inequality for norm of random variables).

□

D Optimization Programs

In this section we describe the optimization programs that we investigate in this work: (1) the oracle optimization program, which is directly tied with the accuracy of estimating the optimal value function on the empirical MDP, (2) the algorithm optimization program, which can be solved using the empirical quantities and finally (3) the $\star$-optimization program, function of problem dependent quantities which can be used to analyze the sample complexity of our algorithm.

D.1 Definitions

Definition 4 (Oracle Minimax Program)**.**

$$\begin{aligned} &\min_n f_{\mathcal{O}}(n) \\ &\quad \textit{s.t.} \quad \sum_{(s,a)} n_{sa} \le n_{max} \end{aligned} \tag{43}$$

where:

$$\begin{aligned} f_{\mathcal{O}}(n) &\stackrel{def}{=} \max_{\rho,\pi} \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k+1,\rho} CI_{sa}(n_{sa}^{k+1}) \\ \textit{s.t.} \quad & (I - \gamma(\widehat{P}_{k+1}^{\pi})^\top)\widehat{\overline{w}}^{\pi,k+1,\rho} = \rho \\ & \sum_s \rho_s = 1 \\ & \rho_{sa} \ge 0 \\ & (\widehat{V}_k^\star - \widehat{V}_k^\pi)(\rho) \le C\epsilon_k \end{aligned} \tag{44}$$

Definition 5 (Algorithm Minimax Program)**.**

$$\begin{aligned} &\min_n f_{\mathcal{A}}(n) \\ &\quad \textit{s.t.} \quad \sum_{(s,a)} n_{sa} \le n_{max} \end{aligned} \tag{45}$$

where:

$$\begin{aligned} f_{\mathcal{A}}(n) &\stackrel{def}{=} \max_{\rho,\pi} \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} (\widehat{CI}_{sa}(n_{sa}^{k+1}) + B_{ksa}) + 2Cp(n_{min})\epsilon_k^\pi \\ \textit{s.t.} \quad & (I - \gamma(\widehat{P}_{k}^{\pi})^\top)\widehat{\overline{w}}^{\pi,k,\rho} = \rho \\ & \sum_s \rho_s = 1 \\ & \rho_{sa} \ge 0 \\ & (\widehat{V}_k^\star - \widehat{V}_k^\pi)(\rho) \le C\epsilon_k \end{aligned} \tag{46}$$

The program is solved with $\epsilon_k^\pi = \epsilon_k$.

Definition 6 ($\star$-Minimax Program)**.**

$$\begin{aligned} &\min_n f_{\star}(n) \\ &\quad \textit{s.t.} \quad \sum_{(s,a)} n_{sa} \le n_{max} \end{aligned} \tag{47}$$

where:

$$\begin{aligned} f_{\star}(n) &\stackrel{def}{=} \max_{\rho,\pi} \sum_{(s,a)} \overline{w}_{sa}^{\pi,\rho} \left(CI_{sa}(n_{sa}^{k+1}) + 2B_{ksa}\right) + 15Cp(n_{min})\epsilon_k^\pi + 8Cp(n_{min})\epsilon_k \\ \textit{s.t.} \quad & (I - \gamma(P^{\pi})^\top)\overline{w}^{\pi,\rho} = \rho \\ & \sum_s \rho_s = 1 \\ & \rho_s \ge 0 \\ & (V^\star - V^\pi)(\rho) \le C^\star\epsilon_k \end{aligned} \tag{48}$$

Similarly as above, $\epsilon_k^\pi = \epsilon_k$ when computing the sample complexity because the program of definition 5 is solved with $\epsilon_k^\pi = \epsilon_k$.

D.2 Relation Between the Optimization Programs

In this section we investigate the relation between the three optimization programs (oracle, algorithm and $\star$). In particular, we show that we can upper bound the objective function of the inner *maximization* program and enlarge its feasibility set as we move from the oracle to the algorithm and finally to the $\star$ program. This ensures that the outer *minimization* is minimizing a function that is increasingly larger (when moving from the oracle to the algorithm and finally to the $\star$ program), giving an upper bound on its value.

Lemma 6 (Relation Between the Optimization Programs)**.** *Consider the three optimization programs of definition 4,5,6. We have that:*

$$f_{\mathcal{O}}(n) \leq f_{\mathcal{A}}(n) \tag{49}$$

Furthermore, outside of the failure event if

$$|(V^\star - \widehat{V}_k^\star)(\rho)| \leq \epsilon_k, \quad \forall \rho \geq 0, \|\rho\|_1 = 1 \tag{50}$$

holds then

$$f_{\mathcal{A}}(n) \leq f_\star(n) \tag{51}$$

$$\tag{52}$$

also holds.

Proof.

Oracle Minimax to Algorithm Minimax Consider the maximization program contained in the definition of $f_{\mathcal{O}}$, see definition 4. First, we can add the variable $\widehat{\overline{w}}^{\pi,k,\rho}$ and the constraint

$$(I - \gamma(\widehat{P}_k^\pi)^\top)\widehat{\overline{w}}^{\pi,k,\rho} = \rho \tag{53}$$

to the oracle inner maximization program without changing its objective value or restricting its feasibility set since $\widehat{\overline{w}}^{\pi,k,\rho}$ is fully determined by equation 53. Next, lemma 9 allows us to move from using the distribution from episode $k+1$ (which is unknown) to episode k (which can be computed by using the empirical MDP $\widehat{\mathcal{M}}_k$) in the objective function:

$$\sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k+1,\rho} CI_{sa}(n_{sa}^{k+1}) = \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} CI_{sa}(n_{sa}^{k+1}) + \sum_{(s,a)} (\widehat{\overline{w}}_{sa}^{\pi,k+1,\rho} - \widehat{\overline{w}}_{sa}^{\pi,k,\rho}) CI_{sa}(n_{sa}^{k+1}) \tag{54}$$

$$\leq \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} CI_{sa}(n_{sa}^{k+1}) + 2Cp(n_{min})\epsilon_k^\pi \tag{55}$$

At this point we can use the variance correction provided by lemma 11 to obtain:

$$\sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} CI_{sa}(n_{sa}^{k+1}) + 2Cp(n_{min})\epsilon_k^\pi \leq \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} (\widehat{CI}_{sa}(n_{sa}^{k+1}) + B_{ksa}) + 2Cp(n_{min})\epsilon_k^\pi \tag{56}$$

that uses the empirical quantities only. We can now drop the variable $\widehat{\overline{w}}^{\pi,k+1,\rho}$ and its constraint

$$(I - \gamma(\widehat{P}_{k+1}^\pi)^\top)\widehat{\overline{w}}^{\pi,k+1,\rho} = \rho \tag{57}$$

since the variable $\widehat{\overline{w}}^{\pi,k+1,\rho}$ does not appears elsewhere. Notice that this again does not change the feasible set for (π, ρ).

Algorithm Minimax to $\star$-Minimax First, we add the variable $\overline{w}^{\pi,\rho}$ and the constraint

$$(I-\gamma(P^\pi)^\top)\overline{w}^{\pi,\rho} = \rho \tag{58}$$

to the minimax optimization program of definition 5; this does not change the objective function or restrict the set of feasible ρ, π since equation 58 can always be satisfied: $\overline{w}^{\pi,\rho}$ is the distribution induced by the policy upon starting from ρ on $\mathcal{M}$. Next, we chain lemma 9 with lemma 11 to express the objective function as a function of the "real quantities", obtaining:

$$\sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} \widehat{CI}_{sa}(n_{sa}^{k+1}) \le \sum_{(s,a)} \overline{w}_{sa}^{\pi,\rho}\left(CI_{sa}(n_{sa}^{k+1}) + B_{ksa}\right) + Cp(n_{min})\epsilon_k^\pi. \tag{59}$$

Now we examine the remaining term:

$$\sum_{(s,a)}\left(\overline{w}_{sa}^{\pi,\rho} - \widehat{\overline{w}}_{sa}^{\pi,k,\rho}\right)(2B_{ksa}) = 2\sum_{(s,a)}\left(\overline{w}_{sa}^{\pi,\rho} - \widehat{\overline{w}}_{sa}^{\pi,k,\rho}\right)\left(\frac{2c_n}{(1-\gamma)(n_{sa}-1)} + \gamma\sqrt{\frac{2c_n}{n_{sa}}}\epsilon_k\right). \tag{60}$$

We further split the above into two. For the first, using the definition of $CI_{sa}(n_{sa}^k)$ and lemma 10 we obtain:

$$2\sum_{(s,a)}\left(\overline{w}_{sa}^{\pi,\rho} - \widehat{\overline{w}}_{sa}^{\pi,k,\rho}\right)\left(\frac{2c_n}{(1-\gamma)(n_{sa}-1)}\right) = 2\sum_{(s,a)}\left(\overline{w}_{sa}^{\pi,\rho} - \widehat{\overline{w}}_{sa}^{\pi,k,\rho}\right)\left(6 \times \underbrace{\frac{c_n}{3(1-\gamma)(n_{sa}-1)}}_{CI_{sa}(n_{sa}^k)}\right) \tag{61}$$

$$\le 12\sum_{(s,a)}\left(\overline{w}_{sa}^{\pi,\rho} - \widehat{\overline{w}}_{sa}^{\pi,k,\rho}\right) CI_{sa}(n_{sa}^k) \le 12Cp(n_{min})\epsilon_k^\pi. \tag{62}$$

For the second term, use the definition of $Cp(n_{min})$ together with $\sum_{(s,a)}\left(\overline{w}_{sa}^{\pi,\rho} + \widehat{\overline{w}}_{sa}^{\pi,k,\rho}\right) = \frac{2}{(1-\gamma)}$ to claim:

$$2\sum_{(s,a)}\left(\overline{w}_{sa}^{\pi,\rho} - \widehat{\overline{w}}_{sa}^{\pi,k,\rho}\right)\left(\gamma\sqrt{\frac{2c_n}{n_{sa}}}\epsilon_k\right) \le 4\sum_{(s,a)}\left(\overline{w}_{sa}^{\pi,\rho} + \widehat{\overline{w}}_{sa}^{\pi,k,\rho}\right)(1-\gamma)Cp(n_{min})\epsilon_k = 8Cp(n_{min})\epsilon_k. \tag{63}$$

In summary we have obtained:

$$\sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho}\left(\widehat{CI}_{sa}(n_{sa}^{k+1}) + B_{ksa}\right) + 2Cp(n_{min}) \tag{64}$$

$$\le \sum_{(s,a)} \overline{w}_{sa}^{\pi,\rho}\left(CI_{sa}(n_{sa}^{k+1}) + 2B_{ksa}\right) + 15Cp(n_{min})\epsilon_k^\pi + 8Cp(n_{min})\epsilon_k. \tag{65}$$

At this point we can drop the the variable $\widehat{\overline{w}}^{\pi,k,\rho}$ and its constraint:

$$(I-\gamma(\widehat{P}_{k+1}^\pi)^\top)\widehat{\overline{w}}^{\pi,k,\rho} = \rho \tag{66}$$

since the variable $\widehat{\overline{w}}^{\pi,k,\rho}$ does not show up elsewhere. Notice that this operation does not change the constraints on the ρ, π optimization variables. Finally, lemma 7 allows us to replace the constraint $(\widehat{V}_k^\star - \widehat{V}_k^\pi)(\rho) \le C\epsilon_k$ with the relaxed version on the real MDP $(V^\star - V^\pi)(\rho) \le C^\star\epsilon_k$. This enlarges the feasibility set for the policies. Notice that an enlarged feasibility for the maximization program set can only increase the objective function, so $f_{\mathcal{O}} \le f_\star$ holds pointwise. $\square$

E Helper Lemmas

In this section we state and prove some helper lemmas.

E.1 Enlarging The Feasibility Set

Lemma 7 (Enlarging The Feasibility Set)**.** *Outside of the failure event if:*

$$|(V^\star - \widehat{V}^\star_k)(\rho)| \le \epsilon_k, \quad \forall \rho \ge 0, \|\rho\|_1 = 1 \tag{67}$$

$$\pi \in \widehat{S}_k \tag{68}$$

$$\epsilon^\pi_k \le \epsilon_k \tag{69}$$

then

$$\pi \in S_k. \tag{70}$$

Proof. Since by assumption:

$$(\widehat{V}^\star_k - \widehat{V}^\pi_k)(\rho) \le C\epsilon_k \tag{71}$$

for all ρ then must we have that (by choosing ρ to be any canonical vector in $\mathbb{R}^{|\mathcal{S}|}$):

$$\|\widehat{V}^\star_k - \widehat{V}^\pi_k\|_\infty \le C\epsilon_k. \tag{72}$$

Next, for an arbitrary ρ consider:

$$(\widehat{V}^\star_k - \widehat{V}^\pi_k)(\rho) = (\underbrace{\widehat{V}^\star_k - V^\star}_{\ge -\epsilon_k} + V^\star - V^\pi + V^\pi - \widehat{V}^\pi_k)(\rho) \tag{73}$$

and apply the simulation lemma, lemma 4 to the last difference obtaining:

$$\ge -\epsilon_k + (V^\star - V^\pi)(\rho) - \epsilon_k - Cp(n_{min})\|V^\star - V^\pi\|_\infty \tag{74}$$

Therefore:

$$-2\epsilon_k + (V^\star - V^\pi)(\rho) - Cp(n_{min})\|V^\star - V^\pi\|_\infty \le (\widehat{V}^\star_k - \widehat{V}^\pi_k)(\rho) \le C\epsilon_k \tag{75}$$

from which we can derive:

$$(V^\star - V^\pi)(\rho) \le C\epsilon_k + 2\epsilon_k + Cp(n_{min})\|V^\star - V^\pi\|_\infty. \tag{76}$$

By taking ρ to be each canonical vector in $\mathbb{R}^{\mathcal{S}}$ we can derive:

$$\|V^\star - V^\pi\|_\infty \le C\epsilon_k + 2\epsilon_k + Cp(n_{min})\|V^\star - V^\pi\|_\infty. \tag{77}$$

from which

$$\|V^\star - V^\pi\|_\infty \le \frac{C\epsilon_k + 2\epsilon_k}{1 - Cp(n_{min})}. \tag{78}$$

follows.[8] Since this is a max-norm bound on the vector $V^\star - V^\pi$, a linear combination weighted by ρ must also satisfy:

$$(V^\star - V^\pi)(\rho) \le \frac{C+2}{1 - Cp(n_{min})}\epsilon_k \stackrel{def}{=} C^\star\epsilon_k. \tag{79}$$

The thesis finally follows by definition of S_k. □

[8] For this passage we need $Cp(n_{min}) < 1$, but our choice of n_{min} ensures that.

E.2 Visit Probability Lemma

Although the visit distribution $\overline{w}^{\pi,\rho}$ can require many samples to be estimated accurately, in this section we show that the uncertainty in the distribution nicely interacts with the confidence intervals for the transitions and rewards. We need the following helper lemma:

Lemma 8 (Distribution Lemma)**.** *It holds that:*

$$\left(\overline{w}^{\pi,\rho} - \widehat{\overline{w}}^{\pi,k,\rho}\right)^\top = \left(\overline{w}^{\pi,\rho}\right)^\top \left(P^\pi - \widehat{P}^\pi_k\right) \sum_{t=0}^{\infty} \gamma^{t+1} \left(\widehat{P}^\pi_k\right)^t. \tag{80}$$

Proof. The cumulative discounted sum of visit probabilities (for policy π upon starting from ρ) satisfies (see for example [WBS07])

$$\overline{w}^{\pi,\rho} = \rho + \gamma (P^\pi)^\top \overline{w}^{\pi,\rho} \tag{81}$$

$$\widehat{\overline{w}}^{\pi,k,\rho} = \rho + \gamma (\widehat{P}^\pi)^\top \widehat{\overline{w}}^{\pi,k,\rho} \tag{82}$$

on $\mathcal{M}$ and $\widehat{\mathcal{M}}_k$, respectively. Subtraction yields:

$$\begin{aligned}
\overline{w}^{\pi,\rho} - \widehat{\overline{w}}^{\pi,k,\rho} &= \gamma \left((P^\pi)^\top \overline{w}^{\pi,\rho} - (\widehat{P}^\pi_k)^\top \widehat{\overline{w}}^{\pi,k,\rho} \right) &(83)\\
&= \gamma \left((P^\pi)^\top \overline{w}^{\pi,\rho} - (\widehat{P}^\pi_k)^\top \overline{w}^{\pi,\rho} + (\widehat{P}^\pi_k)^\top \overline{w}^{\pi,\rho} - (\widehat{P}^\pi_k)^\top \widehat{\overline{w}}^{\pi,k,\rho} \right) &(84)\\
&= \gamma \left((P^\pi - \widehat{P}^\pi_k)^\top \overline{w}^{\pi,\rho} + (\widehat{P}^\pi_k)^\top \left(\overline{w}^{\pi,\rho} - \widehat{\overline{w}}^{\pi,k,\rho} \right) \right). &(85)
\end{aligned}$$

By induction, this yields:

$$\overline{w}^{\pi,\rho} - \widehat{\overline{w}}^{\pi,k,\rho} = \sum_{t=0}^{\infty} \gamma^{t+1} \left((\widehat{P}^\pi_k)^\top \right)^t \left((P^\pi - \widehat{P}^\pi_k)^\top \overline{w}^{\pi,\rho} \right). \tag{86}$$

By transposing the above equality we obtain the statement. □

Now we are ready to analyze how the distribution shift interacts with the confidence intervals for policies that are accurately estimated:

Lemma 9 (Interaction between distribution inaccuracy and confidence intervals)**.** *If for π it holds that*

$$(\widehat{\overline{w}}^{\pi,k,\rho})^\top CI^k \le \epsilon^\pi_k \tag{87}$$

then we have that:

$$\begin{aligned}
\left(\overline{w}^{\pi,\rho} - \widehat{\overline{w}}^{\pi,k,\rho}\right)^\top CI^k &\le Cp(n_{min}) \epsilon^\pi_k \\
\left(\widehat{\overline{w}}^{\pi,j,\rho} - \widehat{\overline{w}}^{\pi,k,\rho}\right)^\top CI^k &\le 2Cp(n_{min}) \epsilon^\pi_k, \quad j \ge k. &(88)
\end{aligned}$$

Proof. Thanks to lemma 8 we can write:

$$\begin{aligned}
\left(\overline{w}^{\pi,\rho} - \widehat{\overline{w}}^{\pi,k,\rho}\right)^\top CI^k &= \gamma \left(\overline{w}^{\pi,\rho}\right)^\top \left(P^\pi - \widehat{P}^\pi_k\right) \sum_{t=0}^{\infty} \gamma^t \left(\widehat{P}^\pi_k\right)^t CI^k &(89)\\
&\le \gamma \| \left(\overline{w}^{\pi,\rho}\right)^\top \left(P^\pi - \widehat{P}^\pi_k\right) \|_1 \| \sum_{t=0}^{\infty} \gamma^t \left(\widehat{P}^\pi_k\right)^t CI^k \|_\infty &(90)
\end{aligned}$$

Notice that the j-th row of

$$\sum_{t=0}^{\infty} \gamma^t \left(\widehat{P}^\pi_k\right)^t \tag{91}$$

is precisely the discounted sum of visit probabilities upon starting from state j on $\widehat{\mathcal{M}}_k$ and following π. Let us call e_j the canonical vector with a 1 in position j and 0's elsewhere; the j-th row is expressible as:

$$e_j^\top \sum_{t=0}^{\infty} \gamma^t \left(\widehat{P}_k^\pi\right)^t = (\widehat{\overline{w}}^{\pi,k,s_j})^\top. \tag{92}$$

This immediately yields:

$$e_j^\top \sum_{t=0}^{\infty} \gamma^t \left(\widehat{P}_k^\pi\right)^t CI^k = (\widehat{\overline{w}}^{\pi,k,s_j})^\top CI^k \leq \epsilon_k^\pi. \tag{93}$$

and therefore equation 89 admits the upper bound (by Holder's inequality):

$$\begin{aligned}
&\leq \gamma \| \left(\overline{w}^{\pi,\rho}\right)^\top (P^\pi - \widehat{P}_k^\pi)\|_1 \epsilon_k^\pi && (94)\\
&= \gamma \| \sum_{(s,a)} \overline{w}_{sa}^{\pi,\rho} \left(p(s,a) - \widehat{p}_k(s,a)\right)^\top \|_1 \epsilon_k^\pi && (95)\\
&\leq \gamma \sum_{(s,a)} \overline{w}_{sa}^{\pi,\rho} \| \left(p(s,a) - \widehat{p}_k(s,a)\right)^\top \|_1 \epsilon_k^\pi && (96)\\
&\leq Cp(n_{min})\epsilon_k^\pi. && (97)
\end{aligned}$$

To obtain the second equation of 88 proceed similarly:

$$\begin{aligned}
\left(\widehat{\overline{w}}^{\pi,j,\rho} - \widehat{\overline{w}}^{\pi,k,\rho}\right)^\top CI^k &= \gamma \left(\widehat{\overline{w}}^{\pi,j,\rho}\right)^\top (\widehat{P}_j^\pi - \widehat{P}_k^\pi) \sum_{t=0}^{\infty} \gamma^t \left(\widehat{P}_k^\pi\right)^t CI^k && (98)\\
&\leq \gamma \| \left(\widehat{\overline{w}}^{\pi,j,\rho}\right)^\top (\widehat{P}_j^\pi - \widehat{P}_k^\pi)\|_1 \| \sum_{t=0}^{\infty} \gamma^t \left(\widehat{P}_k^\pi\right)^t CI^k \|_\infty && (99)\\
&\leq \gamma \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,j,\rho} \left(\|\widehat{p}_j(s,a) - p(s,a)\|_1 + \|p(s,a) - \widehat{p}_k(s,a)\|_1\right) \| \sum_{t=0}^{\infty} \gamma^t \left(\widehat{P}_k^\pi\right)^t CI^k \|_\infty && (100)\\
&\leq 2Cp(n_{min})\epsilon_k^\pi. && (101)
\end{aligned}$$

□

E.3 Bernstein Correction

In this section we discuss how to correct the variance estimate computed with the empirical transitions and value function estimate in a way that results in a variance overestimate (to obtain valid confidence intervals).

Lemma 10 (Bernstein Correction)**.** *If*

$$\|V^\star - \widehat{V}_k^\star\|_\infty \leq \epsilon_k \tag{102}$$

holds then outside of the failure event we have that:

$$\Big| \sqrt{\operatorname*{Var}_{p(s,a)} V^\star} - \sqrt{\operatorname*{Var}_{\widehat{p}_k(s,a)} \widehat{V}_k^\star} \Big| \leq \epsilon_k + \frac{1}{1-\gamma}\sqrt{\frac{2c_n}{n_{sa}-1}} \tag{103}$$

Proof. Outside the failure event:

$$\Big| \sqrt{\operatorname*{Var}_{\widehat{p}_k(s,a)} V^\star} - \sqrt{\operatorname*{Var}_{p(s,a)} V^\star} \Big| \leq \frac{1}{1-\gamma}\sqrt{\frac{2c_n}{n_{sa}-1}} \tag{104}$$

holds and further lemma 5 yields:

$$\Big| \sqrt{\operatorname*{Var}_{\widehat{p}_k(s,a)} V^\star} - \sqrt{\operatorname*{Var}_{\widehat{p}_k(s,a)} \widehat{V}_k^\star} \Big| \leq \|V^\star - \widehat{V}_k^\star\|_\infty. \tag{105}$$

Chaining the two yields the statement. □

Lemma 11 (Realation Between Real and Empirical Confidence Intervals). *If*

$$\|V^\star - \widehat{V}_k^\star\|_\infty \le \epsilon_k \tag{106}$$

holds then outside of the failure event we have that:

$$CI_{sa}(n_{sa}) \le \widehat{CI}_{sa}(n_{sa}) + B_{ksa} \le CI_{sa}(n_{sa}) + 2B_{ksa} \tag{107}$$

where:

$$B_{ksa} \stackrel{def}{=} \frac{2c_n}{(1-\gamma)(n_{sa}-1)} + \gamma\sqrt{\frac{2c_n}{n_{sa}}}\epsilon_k. \tag{108}$$

Proof. Apply lemma 10 twice, first to obtain the "hat quantities" and then to go back to the "real quantities":

$$CI_{sa}(n_{sa}) \stackrel{def}{=} \sqrt{\frac{2\operatorname{Var} R(s,a)c_n}{n_{sa}}} + \gamma\sqrt{\frac{2\operatorname{Var}_{p(s,a)} V^\star c_n}{n_{sa}}} + \underbrace{\frac{c_n}{3(n_{sa}-1)} + \frac{\gamma c_n}{3(1-\gamma)(n_{sa}-1)}}_{=\frac{c_n}{3(1-\gamma)(n_{sa}-1)}} \tag{109}$$

$$\le \sqrt{\frac{2\operatorname{Var}\widehat{R}(s,a)c_n}{n_{sa}}} + \gamma\sqrt{\frac{2\operatorname{Var}_{\widehat{p}_k(s,a)} \widehat{V}_k^\star c_n}{n_{sa}}} + \frac{c_n}{3(1-\gamma)(n_{sa}-1)} + \underbrace{\frac{2c_n}{(1-\gamma)(n_{sa}-1)} + \gamma\sqrt{\frac{2c_n}{n_{sa}}}\epsilon_k}_{B_{ksa}} \tag{110}$$

$$\stackrel{def}{=} \widehat{CI}_{sa}(n_{sa}) + B_{ksa} \tag{111}$$

$$\le \sqrt{\frac{2\operatorname{Var} R(s,a)c_n}{n_{sa}}} + \gamma\sqrt{\frac{2\operatorname{Var}_{p(s,a)} V^\star c_n}{n_{sa}}} + \frac{c_n}{3(1-\gamma)(n_{sa}-1)} + \underbrace{2\left(\frac{2c_n}{(1-\gamma)(n_{sa}-1)} + \gamma\sqrt{\frac{2c_n}{n_{sa}}}\epsilon_k\right)}_{2B_{ksa}} \tag{112}$$

$$\stackrel{def}{=} CI_{sa}(n_{sa}) + 2B_{ksa}. \tag{113}$$

□

E.4 Feasible Set Contains Good Policies

In this section we build the supporting lemmas to show that the feasibility set $(\widehat{V}_k^\star - \widehat{V}_k^\pi)(\rho) \le C\epsilon_k$ in episode k is constructed in a way that ensures that the optimal policy $\pi^\star$ and the next-episode empirical optimal policy $\widehat{\pi}_{k+1}^\star$ are never eliminated at step k, i.e., they are $C\epsilon_k$-optimal for all starting distribution ρ. This ensures that enough samples are allocated at step k to use lemma 2 at the next episode. This guarantees an accurate estimate of the value function.

First we focus on the optimal policy $\pi^\star$.

Lemma 12 ($\pi^\star$ is a Feasible Solution). *Outside of the failure event, if*

$$\|\widehat{V}_k^\star - V^\star\|_\infty \le \epsilon_k \tag{114}$$

$$\epsilon_k^{\pi^\star} \le \epsilon_k \tag{115}$$

holds at step k then it holds that:

$$(\widehat{V}_k^\star - \widehat{V}_k^{\pi^\star})(\rho) \le 2\epsilon_k \le C\epsilon_k, \quad \forall\rho \ge 0, \|\rho\|_1 = 1 \tag{116}$$

Proof. Using the simulation lemma (lemma 3) with $\pi = \pi^\star$ and using the fact that we are outside of the failure event we obtain that:

$$|(\widehat{V}_k^\star - V^\star + V^\star - \widehat{V}_k^{\pi^\star})(\rho)| \le |(\widehat{V}_k^\star - V^\star)(\rho)| + |(V^\star - \widehat{V}_k^{\pi^\star})(\rho)| \tag{117}$$

$$\le \epsilon_k + \Big|\sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi^\star,k,\rho} CI_{sa}(n_{sa}^k)\Big| \le \epsilon_k + \epsilon_k^{\pi^\star} \le 2\epsilon_k. \tag{118}$$

□

Next we turn our attention to the next-step empirical optimal policy. To ensure accuracy, we need to show that we always allocate enough samples to $\widehat{\pi}^\star_{k+1}$ at step k, i.e., $\widehat{\pi}^\star_{k+1}$ is feasible at step k. We achieve this through an inductive argument in theorem 1, which leverages the following lemma. The lemma shows that if a policy is ruled out then it can never become optimal again. This lemma plays a key role in constructing the constraint $(\widehat{V}^\star_k - \widehat{V}^\pi_k)(\rho) \le C\epsilon_k$ because it defines its size through the constant C.

Lemma 13 (Ruled-Out Policies Can Never Be Optimal Again)**.** *Let j be the first episode in which policy μ is not feasible for some ρ in the sense that $\mu \notin \widehat{S}_j$ while $\mu \in \widehat{S}_{j-1}, \mu \in \widehat{S}_{j-2}, \dots, \mu \in \widehat{S}_1$ holds. Outside of the failure event if $\pi^\star \in \widehat{S}_j, C \ge 20, \epsilon_j = \epsilon_{j-1}/2$ and:*

$$|(V^\star - \widehat{V}^\star_j)(\rho)| \le \epsilon_j, \quad \forall \rho \ge 0, \|\rho\|_1 = 1 \tag{119}$$

$$|(V^\star - \widehat{V}^\star_{j-1})(\rho)| \le \epsilon_{j-1}, \quad \forall \rho \ge 0, \|\rho\|_1 = 1 \tag{120}$$

hold together with:

$$2\left(\epsilon^\mu_{j-1} + 2Cp(n_{min})\epsilon^\mu_{j-1} + Cp(n_{min})(1+C)\epsilon_{j-1}\right) \le 4\epsilon_{j-1} \tag{121}$$

$$2\left(\epsilon^{\widehat{\pi}^\star_j}_j + 2Cp(n_{min})\epsilon^{\widehat{\pi}^\star_j}_j + Cp(n_{min})(1+C)\epsilon_j\right) \le 4\epsilon_j \tag{122}$$

then μ cannot be an optimal policy on any empirical MDP $\widehat{\mathcal{M}}_k$ for $k \ge j$.

Proof. Coupled with the hypothesis, Lemma 14 ensures:

$$|(\widehat{V}^{\widehat{\pi}^\star_j}_k - \widehat{V}^{\widehat{\pi}^\star_j}_j)(\rho)| \le 4\epsilon_j \tag{123}$$

$$|(\widehat{V}^\mu_k - \widehat{V}^\mu_{j-1})(\rho)| \le 4\epsilon_{j-1} \tag{124}$$

$$|(\widehat{V}^\mu_j - \widehat{V}^\mu_{j-1})(\rho)| \le 4\epsilon_{j-1} \tag{125}$$

Since $\mu \notin \widehat{S}_j$ by assumption, we have that for at least a starting distribution ρ:

$$(\widehat{V}^{\widehat{\pi}^\star_j}_j - \widehat{V}^\mu_j)(\rho) > C\epsilon_j = \frac{1}{2}C\epsilon_{j-1}. \tag{126}$$

This implies that for that starting distribution ρ:

$$\begin{aligned}
(\widehat{V}^\mu_k - \widehat{V}^\star_k)(\rho) &\le (\widehat{V}^\mu_k - \widehat{V}^{\widehat{\pi}^\star_j}_k)(\rho) && (127)\\
&= (\widehat{V}^\mu_k - \widehat{V}^\mu_{j-1} + \widehat{V}^\mu_{j-1} - \widehat{V}^\mu_j + \widehat{V}^\mu_j - \widehat{V}^{\widehat{\pi}^\star_j}_j + \widehat{V}^{\widehat{\pi}^\star_j}_j - \widehat{V}^{\widehat{\pi}^\star_j}_k)(\rho) && (128)\\
&< 8\epsilon_{j-1} + 4\epsilon_j - \frac{1}{2}C\epsilon_{j-1} = 8\epsilon_{j-1} + 2\epsilon_{j-1} - \frac{1}{2}C\epsilon_{j-1} \le 0 && (129)
\end{aligned}$$

In other words μ is not optimal on $\widehat{\mathcal{M}}_k$. □

The following helper lemma explains that the empirical value of a policy doesn't change a lot between different episodes provided that the policy is feasible for the smaller-numbered episode. In other words, if a policy is accurately estimated, say of order ϵ, its value on all empirical MDPs for later episodes has a fluctuation of order ϵ.

Lemma 14 (Empirical Value of Feasible Policies Does Not Change Much In Later Episodes)**.** *If $\pi \in \widehat{S}_j$ and*

$$|(V^\star - \widehat{V}^\star_j)(\rho)| \le \epsilon_j, \quad \forall \rho \ge 0, \|\rho\|_1 = 1 \tag{130}$$

also holds, then outside of the failure event it holds that:

$$|\left(\widehat{V}^\pi_k - \widehat{V}^\pi_j\right)(\rho)| \le (2 + 4Cp(n_{min}))\epsilon^\pi_j + 2Cp(n_{min})(1+C)\epsilon_j \tag{131}$$

for all episodes $k \ge j$.

Proof. The simulation lemma 3 applied to the two empirical MDPs $\widehat{\mathcal{M}}_k$ and $\widehat{\mathcal{M}}_j$ gives:

$$\left(\widehat{V}_k^{\pi} - \widehat{V}_j^{\pi}\right)(\rho) = \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} \left((\widehat{r}_k(s,a) - \widehat{r}_j(s,a)) + \gamma \left(\widehat{p}_k(s,a) - \widehat{p}_j(s,a)\right)^{\top} \widehat{V}_j^{\pi} \right) \quad (132)$$

$$= \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} \left((\widehat{r}_k(s,a) - \widehat{r}_j(s,a)) + \gamma \left(\widehat{p}_k(s,a) - \widehat{p}_j(s,a)\right)^{\top} V^{\star} \right) \quad (133)$$

$$+ \gamma \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} \left(\widehat{p}_k(s,a) - \widehat{p}_j(s,a)\right)^{\top} (\widehat{V}_j^{\pi} - V^{\star}) \quad (134)$$

We focus on the first term. Outside the failure event the first upper bound below holds

$$\leq \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} \left(CI_{sa}(n_{sa}^k) + CI_{sa}(n_{sa}^j) \right) \leq 2 \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} CI_{sa}(n_{sa}^j). \quad (135)$$

The second upper bound above holds because $n_{sa}^k \geq n_{sa}^j$ (i.e., number of samples can only increase) and the confidence intervals are shrinking with increasing samples: $CI_{sa}(n_{sa}^k) \leq CI_{sa}(n_{sa}^j)$. Lemma 9 allows[9] us to use the empirical distribution from step j instead of k ensuring:

$$2 \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} CI_{sa}(n_{sa}^j) \leq 2 \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,j,\rho} CI_{sa}(n_{sa}^j) + 4Cp(n_{min})\epsilon_j^{\pi} \quad (136)$$

$$\leq (2 + 4Cp(n_{min}))\epsilon_j^{\pi}. \quad (137)$$

The second step is by definition of ϵ_j^{π}. Now we bound the remaining term; by the hypothesis of this lemma:

$$|(\widehat{V}_j^{\pi} - V^{\star})(\rho)| = |(\widehat{V}_j^{\pi} - \widehat{V}_j^{\star} + \widehat{V}_j^{\star} - V^{\star})(\rho)| \leq (C+1)\epsilon_j \quad (138)$$

for all ρ and so in particular we must have $\|\widehat{V}_j^{\pi} - V^{\star}\|_{\infty} \leq (1+C)\epsilon_j$ and hence outside of the failure event by Holder (and by adding and subtracting $p(s,a)$) it holds that:

$$\gamma \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} \left(\widehat{p}_k(s,a) - \widehat{p}_j(s,a)\right)^{\top} (\widehat{V}_j^{\star} - V^{\star}) \leq 2Cp(n_{min})(1+C)\epsilon_j \quad (139)$$

giving the thesis. □

E.5 Feasible Policies Will Have Improved Accuracy

In this section we show that feasible policies for the algorithm minimax program of definition 5 will gain accuracy at the next episode provided that they are already accurately estimated in the current episode (condition $\epsilon_k^{\pi} \leq \epsilon_k$, which is equivalent to a small distribution shift).

Lemma 15 (Feasible Policies Will Have Improved Accuracy)**.** *Outside of the failure event if:*

$$\pi \in \widehat{S}_k; \quad \epsilon_k^{\pi} \leq \epsilon_k \quad (140)$$

holds then at the next episode it holds that:

$$\epsilon_{k+1}^{\pi} \leq \epsilon_{k+1} = \frac{\epsilon_k}{2}. \quad (141)$$

Proof. BESPOKE solves a program equivalent to the minimax program of definition 5 that ensures that for feasible (π, ρ):

$$\frac{\epsilon_k}{4} \leq \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} \left(\widehat{CI}_{sa}(n_{sa}^k) + B_{ksa} \right) + 2Cp(n_{min})\epsilon_k \leq \frac{\epsilon_k}{2}. \quad (142)$$

by the choice of n_{max} (see inner loop over n_{max} in the algorithm 1). Since $\epsilon_k^{\pi} \leq \epsilon_k$, we have that:

$$\frac{\epsilon_k}{4} \leq \sum_{(s,a)} \widehat{\overline{w}}_{sa}^{\pi,k,\rho} \left(\widehat{CI}_{sa}(n_{sa}^k) + B_{ksa} \right) + 2Cp(n_{min})\epsilon_k^{\pi} \leq \frac{\epsilon_k}{2} \quad (143)$$

[9] Note that j and k are flipped in the two lemmas.

must hold. By lemma 6 we have that the value of the oracle objective of definition 4 is less than $\epsilon_k/2$, i.e.:

$$\epsilon^{\pi}_{k+1} \stackrel{def}{=} \sum_{(s,a)} \widehat{w}^{\pi,k+1,\rho}_{sa} CI_{sa}(n^{k+1}_{sa}) \leq \frac{\epsilon_k}{2} \stackrel{def}{=} \epsilon_{k+1} \tag{144}$$

must hold for those (π, ρ). ☐

Lemma 16 (Minimum Number of Samples)**.** *Outside of the failure event if $C = 20$ and*

$$n_{sa} \geq n_{min} \stackrel{def}{=} \frac{2 \times 625^2 \gamma^2 S c_n}{(1-\gamma)^2} \tag{145}$$

then it holds that:

$$Cp(n_{min}) \leq \frac{1}{625} \tag{146}$$

$$2Cp(n_{min})(2 + C) \leq \frac{1}{2} \tag{147}$$

Proof. Immediate, by definition of $Cp(n_{min})$ and n_{min}, see appendix A. ☐

Finally, the following technical lemma ensures that if the optimal policy and empirical optimal policy of step k are feasible in all episodes up to k then the accuracy in value function estimate can be guaranteed.

Lemma 17 (Guaranteed Accuracy)**.** *Outside of the failure event if:*

1. *$\pi^\star \in \widehat{S}_j$ in all episodes $j < k$*

2. *$\widehat{\pi}^\star_k \in \widehat{S}_j$ in all episodes $j < k$*

then

$$|(V^\star - \widehat{V}^\star_k)(\rho)| \leq \epsilon_k, \quad \forall \rho \geq 0, \|\rho\|_1 = 1 \tag{148}$$

holds.

Proof. Chaining lemma 15 for $\pi^\star$ and all episodes up to $k - 1$ gives:

$$\max_{\rho} \sum_{(s,a)} \widehat{w}^{\pi^\star,k,\rho}_{sa} CI_{sa}(n^k_{sa}) \stackrel{def}{=} \epsilon^{\pi^\star}_k \leq \epsilon_k. \tag{149}$$

Likewise, chaining lemma 15 for $\widehat{\pi}^\star_k$ and all episodes up to $k - 1$ gives:

$$\max_{\rho} \sum_{(s,a)} \widehat{w}^{\widehat{\pi}^\star_k,k,\rho}_{sa} CI_{sa}(n^k_{sa}) \stackrel{def}{=} \epsilon^{\widehat{\pi}^\star_k}_k \leq \epsilon_k. \tag{150}$$

Finally, lemma 2 gives the thesis. ☐

F Main Results

In this section we show that the algorithm works as intended, namely it terminates in finite time after logarithmically many iterations and returns with high probability an ϵ_{Input}-correct value function estimate and an almost ϵ_{Input}-suboptimal policy. Finally, we analyze its sample complexity.

F.1 BESPOKE Works as Intended

Theorem 1 (BESPOKE Works as Intended)**.** *With probability at least $1 - \delta$, in every episode k BESPOKE maintains an empirical MDP $\widehat{\mathcal{M}}_k$ such that its optimal value function $\widehat{V}^\star_k$ and its optimal policy $\widehat{\pi}^\star_k$ satisfy:*

$$\|V^\star - \widehat{V}^\star_k\|_\infty \le \epsilon_k, \quad \|V^\star - V^{\widehat{\pi}^\star_k}\|_\infty \le 2\epsilon_k$$

where $\epsilon_{k+1} \overset{def}{=} \frac{\epsilon_k}{2}, \ \forall k$. In particular, when BESPOKE terminates in episode k_{Final} it holds that $\frac{\epsilon_{Input}}{2} \le \epsilon_{k_{Final}} \le \epsilon_{Input}$.

Proof. We reason by induction outside of the failure event, which has measure $1-\delta$ by lemma 2. The inductive hypothesis is over the episodes $k = 1, 2, \dots$ and consists of the following two conditions:

1. $\pi^\star \in \widehat{S}_j$ in all episodes $j < k$
2. $\widehat{\pi}^\star_k \in \widehat{S}_j$ in all episodes $j < k$

In other words, we assume that the optimal policy $\pi^\star$ is feasible up to episode $k-1$, and that the optimal policy on $\widehat{\mathcal{M}}_k$ is also feasible up to episode $k-1$. Before showing the inductive step, notice that the inductive hypothesis together with lemma 17 ensures[10]:

$$\|V^\star - \widehat{V}^\star_p\|_\infty \le \epsilon_p, \quad \forall p < k. \tag{151}$$

Notice that the above equation is equivalent to:

$$|(V^\star - \widehat{V}^\star_p)(\rho)| \le \epsilon_p, \quad \forall \rho \ge 0, \|\rho\|_1 = 1, \quad \forall p < k. \tag{152}$$

by simply choosing ρ to be the point mass in any starting state.

Optimal Policy We have that $\pi^\star \in \widehat{S}_j$ for all episodes $j \le k-1$ by the inductive hypothesis. By repeatedly chaining lemma 15 for all episodes $j = 1, 2, \dots, k$ we can ensure the condition

$$\epsilon_j^{\pi^\star} \le \epsilon_j, \quad \forall j \le k. \tag{153}$$

Lemma 12 then ensures $\pi^\star \in \widehat{S}_k$.

Empirical Optimal Policies We need to show that $\widehat{\pi}^\star_{k+1} \in \widehat{S}_j$ for all $j \le k$. Suppose it isn't, and let us show that this situation cannot arise. Consider the *first* episode $j \le k$ such that $\widehat{\pi}^\star_{k+1} \notin \widehat{S}_j$. Since $\widehat{\pi}^\star_{k+1}$ was in $\widehat{S}_p, \ \forall p < j$, chaining lemma 15 yields:

$$\epsilon_{j-1}^{\widehat{\pi}^\star_{k+1}} \le \epsilon_{j-1}. \tag{154}$$

The inductive hypothesis ensures[11] $\widehat{\pi}^\star_j$ was in $\widehat{S}_p, \ \forall p < j$, and so chaining lemma 15 yields:

$$\epsilon_j^{\widehat{\pi}^\star_j} \le \epsilon_j \tag{155}$$

[10] the lemma ensures $|(V^\star - \widehat{V}^\star_p)(\rho)| \le \epsilon_p$ for all ρ, so choose ρ to be the point mass on a starting state

[11] Notice that this part refers to policy $\widehat{\pi}^\star_j$ and not $\widehat{\pi}^\star_{k+1}$. This condition is ensured by the inductive hypothesis which holds in all episodes up to k.

also holds. Finally, lemma 16 ensures the following two inequalities:

$$2Cp(n_{min})\underbrace{\epsilon_{j-1}^{\widehat{\pi}^{\star}_{k+1}}}_{\leq \epsilon_{j-1}} + 2Cp(n_{min})(1+C)\epsilon_{j-1} \leq \frac{\epsilon_{j-1}}{2}$$

$$2Cp(n_{min})\underbrace{\epsilon_{j}^{\widehat{\pi}^{\star}_{j}}}_{\leq \epsilon_{j}} + 2Cp(n_{min})(1+C)\epsilon_{j} \leq \frac{\epsilon_{j}}{2}. \quad (156)$$

Together, equation 151, 155, 154, 156 satisfy the assumption of lemma 13 (with $\mu = \widehat{\pi}^{\star}_{k+1}$). This gives a contradiction, because the lemma claims that $\widehat{\pi}^{\star}_{k+1}$ cannot be an optimal policy while being $\notin \widehat{S}_j$ for some $j < k$.

The proof by induction is complete and now lemma 17 ensures

$$|(V^{\star} - \widehat{V}^{\star}_{k})(\rho)| \leq \epsilon_k, \quad \forall \rho \geq 0, \|\rho\|_1 = 1, \quad \forall k. \quad (157)$$

When the termination condition for the algorithm BESPOKE are satisfied in episode k_{Final},

$$|(V^{\star} - \widehat{V}^{\star}_{k_{Final}})(\rho)| \leq \epsilon_{k_{Final}}, \quad \forall \rho \geq 0, \|\rho\|_1 = 1 \quad (158)$$

must hold with

$$\frac{\epsilon_{Input}}{2} \leq \epsilon_{k_{Final}} \leq \epsilon_{Input}. \quad (159)$$

Finally, the triangle inequality gives:

$$\|V^{\star} - V^{\pi_{Final}}\|_{\infty} \leq \|V^{\star} - \widehat{V}^{\star}_{k_{Final}}\|_{\infty} + \|\widehat{V}^{\star}_{k_{Final}} - V^{\pi_{Final}}\|_{\infty} \quad (160)$$

$$\leq \epsilon_{k_{Final}} + \|\widehat{V}^{\star}_{k_{Final}} - V^{\pi_{Final}}\|_{\infty} \quad (161)$$

In addition, lemma 4 ensures that for π_{Final} at that episode:

$$\|\widehat{V}^{\star}_{k_{Final}} - V^{\pi_{Final}}\|_{\infty} \leq \epsilon_{Input} + \gamma Cp(n_{min})\|V^{\star} - V^{\pi_{Final}}\|_{\infty} \quad (162)$$

Combining the two gives:

$$\|V^{\star} - V^{\pi_{Final}}\|_{\infty} \leq \frac{2\epsilon_{Input}}{1 - \gamma Cp(n_{min})} \leq 2.03\epsilon_{Input} \quad (163)$$

by the choice of n_{min} (see appendix A and lemma 16). □

F.2 Computational Complexity of BESPOKE

Proposition 4 (Computational Complexity of BESPOKE)**.** *Outside of the failure event* BESPOKE *terminates in at most*

$$\log_2\left(\frac{1}{(1-\gamma)\epsilon_{Input}}\right) + 1 \quad (164)$$

episodes. Let n_{sa}^{Final} be the total number of samples allocated by the algorithm at termination given by theorem 2. Then BESPOKE *at termination has solved at most:*

$$\left(\log_2\left(\frac{1}{(1-\gamma)\epsilon_{Input}}\right) + 1\right) \times \log_2\left(\sum_{(s,a)} n_{sa}^{Final}\right) \quad (165)$$

convex minimization programs as defined in definition 7.

Proof. By the halving rule on ϵ_k, BESPOKE must terminate after at most $\log_2(\frac{1}{(1-\gamma)\epsilon_{Input}}) + 1$ episodes; in addition, if n_{sa}^{Final} is the final number sample allocated by the the algorithm to (s, a) then BESPOKE solves at most $\log_2(\sum_{(s,a)} n_{sa}^{Final})$ convex programs as described in definition 7 at each episode. □

F.3 Sample Complexity of BESPOKE

Theorem 2 (Sample Complexity of the Algorithm BESPOKE). *With probability at least $1-\delta$, the total sample complexity of* BESPOKE *up to episode k is upper bounded by $\sum_{(s,a)} n_{sa}$ where n_{sa} is the total number of samples allocated to the (s,a) pair:*

$$n_{sa} = \tilde{O}\Bigg(\min \Bigg\{ \frac{1}{(1-\gamma)^3(\epsilon_k)^2}, \frac{\operatorname{Var} R(s,a) + \gamma^2 \operatorname{Var}_{p(s,a)} V^\star}{(1-\gamma)^2(\epsilon_k)^2} + \frac{1}{(1-\gamma)^2(\epsilon_k)}, \tag{166}$$

$$\frac{\operatorname{Var} R(s,a) + \gamma^2 \operatorname{Var}_{p(s,a)} V^\star}{\Delta_{s,a}^2} + \frac{1}{(1-\gamma)\Delta_{s,a}} \Bigg\} + \frac{\gamma S}{(1-\gamma)^2} \Bigg). \tag{167}$$

Proof. We show that a sample complexity

$$n_{sa} = \tilde{O}\Bigg(\min \Bigg(\underbrace{\frac{1}{(1-\gamma)^3\epsilon_k^2}}_{A}, \underbrace{\frac{\operatorname{Var} R(s,a) + \gamma^2 \operatorname{Var}_{p(s,a)} V^\star}{(1-\gamma)^2\epsilon_k^2} + \frac{1}{(1-\gamma)^2\epsilon_k}}_{B_{sa}}, \tag{168}$$

$$\underbrace{\frac{\operatorname{Var} R(s,a) + \gamma^2 \operatorname{Var}_{p(s,a)} V^\star}{\Delta_{sa}^2} + \frac{1}{(1-\gamma)\Delta_{sa}}}_{C_{sa}} \Bigg) + \frac{\gamma S}{(1-\gamma)^2} \Bigg) \tag{169}$$

suffices to ensure that the value of the star minimax program of definition 6 is $\leq \frac{\epsilon_k}{4}$. By lemma 6 this is an upper bound on the sample complexity of the algorithm minimax program of definition 5 to guarantee that its objective is $\leq \frac{\epsilon_k}{4}$.

Three cases are possible for the $\min$ of equation 168: either the $\min$ of equation 168 is attained by A or B_{sa} or C_{sa}, and we partition the state-action space accordingly into the sets $\mathcal{A}_k, \mathcal{B}_k, \mathcal{C}_k$ corresponding to whether a certain (s,a) pair attains the minimum of equation 168 with the expression A, B_{sa}, C_{sa}, respectively. In other words, we have the partition: $\mathcal{A}_k \cup \mathcal{B}_k \cup \mathcal{C}_k = \mathcal{S} \times \mathcal{A}$ where:

$$\mathcal{A}_k \overset{def}{=} \{(s,a) \mid A = \arg\max(A, B_{sa}, C_{sa})\} \tag{170}$$

$$\mathcal{B}_k \overset{def}{=} \{(s,a) \mid B_{sa} = \arg\max(A, B_{sa}, C_{sa})\} \tag{171}$$

$$\mathcal{C}_k \overset{def}{=} \{(s,a) \mid C_{sa} = \arg\max(A, B_{sa}, C_{sa})\} \tag{172}$$

$$\tag{173}$$

with ties broken arbitrarily.

Therefore it suffices to bound the terms below:

$$f_\star(n) = \sum_{(s,a)\in\mathcal{A}_k} \overline{w}_{sa}^{\pi,\rho}(CI_{sa}(n_{sa}^{k+1}) + 2B_{ksa}) + \sum_{(s,a)\in\mathcal{B}_k} \overline{w}_{sa}^{\pi,\rho}(CI_{sa}(n_{sa}^{k+1}) + 2B_{ksa}) \tag{174}$$

$$+ \sum_{(s,a)\in\mathcal{C}_k} \overline{w}_{sa}^{\pi,\rho}(CI_{sa}(n_{sa}^{k+1}) + 2B_{ksa}) + 23Cp(n_{min})\epsilon_k. \tag{175}$$

separately for all policies π and starting distributions ρ that satisfy:

$$(V^\star - V^\pi)(\rho) \leq C^\star \epsilon_k. \tag{176}$$

Pairs in $\mathcal{A}_k$ First notice that a sample complexity

$$\tilde{O}\left(\frac{1}{(1-\gamma)^2\epsilon_k} + \frac{1}{(1-\gamma)^2}\right) \tag{177}$$

suffices to ensure

$$\sum_{(s,a)\in\mathcal{A}_k} \overline{w}_{sa}^{\pi,\rho}(2B_{ksa}) \leq \frac{\epsilon_k}{200}. \tag{178}$$

By definition of $CI_{sa}(n_{sa}^{k+1})$ we can write:

$$\sum_{(s,a)\in\mathcal{A}_k} \overline{w}_{sa}^{\pi,\rho} CI_{sa}(n_{sa}^{k+1}) = \sum_{(s,a)\in\mathcal{A}_k} \overline{w}_{sa}^{\pi,\rho}\gamma\sqrt{\frac{2\operatorname{Var}_{p(s,a)} V^\star c_n}{n_{sa}}} + \sum_{(s,a)\in\mathcal{A}_k} \overline{w}_{sa}^{\pi,\rho}\frac{\gamma c_n}{3(1-\gamma)(n_{sa}-1)}$$
$$+ \sum_{(s,a)\in\mathcal{A}_k} \overline{w}_{sa}^{\pi,\rho}\sqrt{\frac{2\operatorname{Var} R(s,a) c_n}{n_{sa}}} + \sum_{(s,a)\in\mathcal{A}_k} \overline{w}_{sa}^{\pi,\rho}\frac{c_n}{3(n_{sa}-1)} \tag{179}$$

Since π is $C^\star\epsilon_k$ optimal for every starting distribution it must be $C^\star\epsilon_k$ optimal in the max-norm as well and hence we have the upper bound below thanks to lemma 5:

$$\leq \sum_{(s,a)\in\mathcal{A}_k} \gamma\overline{w}_{sa}^{\pi,\rho}\left(\sqrt{\frac{2\operatorname{Var}_{p(s,a)} V^\pi c_n}{n_{sa}}} + \frac{\sqrt{2c_n}}{\sqrt{n_{sa}}}C^\star\epsilon_k + \frac{c_n}{3(1-\gamma)(n_{sa}-1)}\right) \tag{180}$$
$$+ \sum_{(s,a)\in\mathcal{A}_k} \overline{w}_{sa}^{\pi,\rho}\left(\sqrt{\frac{2\operatorname{Var}_{p(s,a)} R(s,a) c_n}{n_{sa}}} + \frac{c_n}{3(n_{sa}-1)}\right) \tag{181}$$

We focus on the first term of equation 180. Cauchy-Schwartz gives:

$$\leq \frac{\gamma}{\sqrt{1-\gamma}}\sqrt{\frac{2\sum_{(s,a)\in\mathcal{A}_k}\overline{w}_{sa}^{\pi,\rho}\operatorname{Var}_{p(s,a)} V^\pi c_n}{n}} \tag{182}$$

where $n = \min_{(s,a)\in\mathcal{A}_k} n_{sa}$. Thanks to the law of total variance [AMK12] we have that $\sum_{(s,a)\in\mathcal{A}_k}\overline{w}_{sa}^{\pi,\rho}\operatorname{Var}_{p(s,a)} V^\pi$ is at most the variance of the returns upon following policy π on the true MDP, and it is thus bounded by $1/(1-\gamma)^2$. This gives the upper bound:

$$\leq \frac{1}{(1-\gamma)^{\frac{3}{2}}}\sqrt{\frac{2c_n}{n}} \tag{183}$$

At this point a sample complexity:

$$n_{sa} = \tilde{O}\left(\frac{1}{(1-\gamma)^3\epsilon_k^2}\right) \tag{184}$$
$$n_{sa} = \tilde{O}\left(\frac{1}{(1-\gamma)^2}\right) \tag{185}$$
$$n_{sa} = \tilde{O}\left(\frac{1}{(1-\gamma)^2\epsilon_k}\right) \tag{186}$$

respectively, suffices to ensure that each term in equation 180 (i.e., the transition terms) is less than $\frac{\epsilon_k}{200}$. Since $\epsilon_k \leq \frac{1}{1-\gamma}$ we have that:

$$n_{sa} = \tilde{O}\left(\frac{1}{(1-\gamma)^3\epsilon_k^2} + \frac{1}{(1-\gamma)^2}\right) \tag{187}$$

suffices. Now we focus on the remaining terms (the reward terms), equation 181. We have the upper bound below:

$$\sum_{(s,a)\in\mathcal{A}_k} \overline{w}_{sa}^{\pi,\rho}\left(\sqrt{\frac{2c_n}{n_{sa}}} + \frac{c_n}{3(n_{sa}-1)}\right) \tag{188}$$

Again, a sample complexity of order:

$$\tilde{O}\left(\frac{1}{(1-\gamma)^2\epsilon_k^2}\right) \tag{189}$$

suffices to ensure that each term in 179 is $\leq \frac{\epsilon_k}{200}$. This implies that expression 181 is $\leq \frac{\epsilon_k}{25}$.

Pairs in $\mathcal{B}_k$ Notice that we have

$$\gamma\sqrt{\frac{2\operatorname{Var}_{p(s,a)} V^\star c_n}{n_{sa}}} \leq (1-\gamma)\frac{\epsilon_k}{100} \tag{190}$$
$$\frac{\gamma c_n}{3(1-\gamma)(n_{sa}-1)} \leq (1-\gamma)\frac{\epsilon_k}{100} \tag{191}$$
$$\sqrt{\frac{2\operatorname{Var} R(s,a) c_n}{n_{sa}}} \leq (1-\gamma)\frac{\epsilon_k}{100} \tag{192}$$
$$\frac{c_n}{3(n_{sa}-1)} \leq (1-\gamma)\frac{\epsilon_k}{100} \tag{193}$$
$$2B_{ksa} \leq (1-\gamma)\frac{\epsilon_k}{100} \tag{194}$$

for

$$n_{sa} = \tilde{O}\left(\frac{\gamma^2 \operatorname{Var}_{p(s,a)} V^\star}{(1-\gamma)^2\epsilon_k^2}\right) \tag{195}$$
$$n_{sa} = \tilde{O}\left(\frac{\gamma}{(1-\gamma)^2\epsilon_k}\right) \tag{196}$$
$$n_{sa} = \tilde{O}\left(\frac{\operatorname{Var} R(s,a)}{(1-\gamma)^2\epsilon_k^2}\right) \tag{197}$$
$$n_{sa} = \tilde{O}\left(\frac{1}{(1-\gamma)\epsilon_k}\right) \tag{198}$$
$$n_{sa} = \tilde{O}\left(\frac{1}{(1-\gamma)^2\epsilon_k} + \frac{\gamma^2}{(1-\gamma)^2}\right) \tag{199}$$

respectively. Summing over the (s,a) pairs with their maximum of type B yields:

$$\sum_{(s,a)\in\mathcal{B}_k} \overline{w}_{sa}^{\pi,\rho}(CI_{sa}(n_{sa}^{k+1}) + 2B_{ksa}) \leq \sum_{(s,a)\in\mathcal{B}_k} \overline{w}_{sa}^{\pi,\rho}(1-\gamma)\frac{\epsilon_k}{20} = \frac{\epsilon_k}{20}. \tag{200}$$

Pairs in $\mathcal{C}_k$ In this case notice that we have

$$\gamma\sqrt{\frac{2\operatorname{Var}_{p(s,a)} V^\star c_n}{n_{sa}}} \leq \frac{\Delta_{sa}}{100C^\star} \tag{201}$$
$$\frac{\gamma c_n}{3(1-\gamma)(n_{sa}-1)} \leq \frac{\Delta_{sa}}{100C^\star} \tag{202}$$
$$\sqrt{\frac{2\operatorname{Var} R(s,a) c_n}{n_{sa}}} \leq \frac{\Delta_{sa}}{100C^\star} \tag{203}$$
$$\frac{c_n}{3(n_{sa}-1)} \leq \frac{\Delta_{sa}}{100C^\star} \tag{204}$$
$$2B_{ksa} \leq \frac{\Delta_{sa}}{100C^\star} \tag{205}$$

for

$$n_{sa} = \tilde{O}\left(\frac{\mathrm{Var}_{p(s,a)} V^\star}{\Delta_{sa}^2}\right) \tag{206}$$

$$n_{sa} = \tilde{O}\left(\frac{1}{(1-\gamma)\Delta_{sa}}\right) \tag{207}$$

$$n_{sa} = \tilde{O}\left(\frac{\mathrm{Var}\, R(s,a)}{\Delta_{sa}^2}\right) \tag{208}$$

$$n_{sa} = \tilde{O}\left(\frac{1}{\Delta_{sa}}\right) \tag{209}$$

$$n_{sa} = \tilde{O}\left(\frac{1}{(1-\gamma)\Delta_{sa}}\right) \tag{210}$$

$$\tag{211}$$

respectively. This ensures[12]

$$CI_{sa}(n_{sa}^{k+1}) + 2B_{ksa} \le \gamma\sqrt{\frac{2\,\mathrm{Var}_{p(s,a)} V^\star c_n}{n_{sa}}} + \frac{\gamma c_n}{3(1-\gamma)(n_{sa}-1)} \tag{212}$$

$$+ \sqrt{\frac{2\,\mathrm{Var}\, R(s,a) c_n}{n_{sa}}} + \frac{c_n}{3(n_{sa}-1)} + \frac{\Delta_{sa}}{100C^\star} \tag{213}$$

$$\le \frac{\Delta_{sa}}{20C^\star} \tag{214}$$

Summing over the (s, a) pairs with their maximum of type C:

$$\sum_{(s,a)\in\mathcal{C}_k} \overline{w}_{sa}^{\pi,\rho}(CI_{sa}(n_{sa}^{k+1}) + 2B_{ksa}) \le \frac{1}{20C^\star}\sum_{(s,a)\in\mathcal{C}_k} \overline{w}_{sa}^{\pi,\rho}\Delta_{sa} = \frac{1}{20C^\star}\left(V^\star - V^\pi\right)(\rho) \le \frac{1}{20}\epsilon_k. \tag{215}$$

The equality arises from lemma 1 and the last inequality on the constraint on the policy for the $\star$-optimization program.

Term $23Cp(n_{min})\epsilon_k$ This can be made $\le \frac{\epsilon_k}{25}$ with lemma 16 by using

$$n_{sa} = \tilde{O}\left(\frac{S}{(1-\gamma)^2}\right) \tag{216}$$

samples.

Concluding remarks Summing all the upper bounds just derived for each term ensures that equation 174 is upper bounded by $\frac{\epsilon_k}{4}$ with a total sample complexity as written in the theorem statement. By lemma 6 this is an upper bound on the sample complexity of the algorithm at step k, and since BESPOKE reaches step k after logarithmically-many episodes (see proposition 4), this is also the total sample complexity up to episode k up to log factors.

□

G Efficient Implementation

In this section we rewrite the minimax optimization program of definition 5 (with $\epsilon_k^\pi = \epsilon_k$) into a convex minimization program that can be efficiently solved. First we directly optimize over the distribution[13] w instead of over the policy π and introduce an appropriate scalar slack variable t. This

[12]Note that $C^\star$ is just a constant.

[13]We drop all subscripts on w in this section for simplicity.

allows us to put the inner maximization in the following matrix form (we neglect the constant term $+2Cp(n_{min})\epsilon$):

$$\begin{aligned} \max_x \quad & c^\top x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0 \end{aligned} \tag{217}$$

with

$$c = \begin{bmatrix} \widehat{CI}^{k+1} + B_k \\ 0 \\ 0 \end{bmatrix}; \quad x = \begin{bmatrix} w \\ \rho \\ t \end{bmatrix}; \quad A = \begin{bmatrix} \Xi - \gamma \widehat{P}_k^\top & -I & 0 \\ 0 & \mathbb{1}^\top & 0 \\ \widehat{r}_k^\top & -\widehat{V}_k^\star & -1 \end{bmatrix}; \quad b = \begin{bmatrix} 0 \\ 1 \\ -C\epsilon_k \end{bmatrix}.$$

Here I is the identity and Ξ is a marginalization matrix as described in [WBS07]. Written explicitly, we have:

$$[\widehat{CI}^{k+1} + B_k]_{sa} = \left(\sqrt{2c_n \operatorname{Var}\left(\widehat{R}(s,a)\right)} + \gamma \sqrt{2c_n \operatorname*{Var}_{\widehat{p}_k(s,a)} \widehat{V}_k^\star} + \gamma \sqrt{2c_n} \epsilon_k \right) \left(\frac{1}{\sqrt{n_{sa}}} \right) \tag{218}$$

$$+ \left(\frac{7c_n}{3(1-\gamma)} \right) \left(\frac{1}{n_{sa} - 1} \right) \tag{219}$$

Notice that the above expression is a convex function of the n_{sa} for $n_{sa} \geq 2$.

We compute the dual of the linear program above (with n_{sa} fixed):

$$\begin{aligned} \min_y \quad & b^\top y \\ \text{s.t.} \quad & A^\top y \geq c \end{aligned} \tag{220}$$

Therefore, the minimax program of definition 5 can be reformulated into an equivalent convex minimization program (now we add back $+2Cp(n_{min})\epsilon$ and the outer minimization program):

Definition 7 (Convex Minimization Program)**.**

$$\begin{aligned} \min_{n,y} \quad & b^\top y + 2Cp(n_{min})\epsilon \\ \textit{s.t.} \quad & c - A^\top y \leq 0 \\ & n_{sa} \geq 0, \quad \forall (s,a) \\ & \sum_{(s,a)} n_{sa} \leq n_{max}. \end{aligned} \tag{221}$$