OPTIMAL APPROXIMATION OF AVERAGE REWARD MARKOV DECISION PROCESSES

Yuri Sapronov

MIPT

Dolgoprudny, Russia

{sapronov.iuf}@phystech.edu

Nikita Yudin

MIPT

Dolgoprudny, Russia

{iudin.ne}@phystech.edu

IITP RAS, ISP RAS, FRC CSC RAS

Moscow, Russia

ABSTRACT

We continue to develop the concept of studying the ε-optimal policy for Average Reward Markov Decision Processes (AMDP) by reducing it to Discounted Markov Decision Processes (DMDP). Existing research often stipulates that the discount factor must not fall below a certain threshold. Typically, this threshold is close to one, and as is well-known, iterative methods used to find the optimal policy for DMDP become less effective as the discount factor approaches this value.

Our work distinguishes itself from existing studies by allowing for inaccuracies in solving the empirical Bellman equation. Despite this, we have managed to maintain the sample complexity that aligns with the latest results. We have succeeded in separating the contributions from the inaccuracy of approximating the transition matrix and the residuals in solving the Bellman equation in the upper estimate so that our findings enable us to determine the total complexity of the epsilon-optimal policy analysis for DMDP across any method with a theoretical foundation in iterative complexity.

1 INTRODUCTION

In recent times, the concept of Reinforcement Learning (RL) has demonstrated exceptional results in various sequential learning and decision-making tasks. These empirical successes have spurred extensive theoretical research into RL algorithms and their fundamental limitations. Typically, the environment in Reinforcement Learning is modeled as a Markov Decision Process (MDP), where the primary objective is to find a policy π that maximizes the expected cumulative reward. There are various criteria for calculating the total reward, such as the finite horizon total reward and the discounted infinite horizon, formally represented as $\mathbb{E}^{\pi}[\sum_{t=0}^{T-1} r_t]$ and $\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$, respectively. Here, r_t is the reward received at the t-th step (see 1.2 for a formal description).

Mathematically, the discount factor is required for the convergence of the series, which allows for comparing the final rewards, and thereby the policies that have achieved them. From a physical standpoint, this approach can be explained by the fact that the most significant steps are the first $\sim \frac{1}{1-\gamma}$ (where $\gamma < 1$). In many practical situations, when long-term policy effectiveness is of interest, we can evaluate the policy in terms of the average accumulated reward: $\lim_{T\to\infty} \mathbb{E}^{\pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} r_t\right]$.

A fundamental theoretical problem in RL is the sample complexity for learning an approximately optimal policy when access is available to what is called a generative model. This implies that we can sample a step for any available state-action pair in the environment, thus interacting with the environment as a "gray box."

As the topic of generative models has gained increasing interest, new approaches for finding an ε-optimal policy for both DMDPs and AMDPs have emerged. In fact, the latter problem can be reduced to the former, which is the central idea behind the approach of reducing AMDP to DMDP, and in this work, we focus on the task of finding a near-optimal policy specifically for the discounted case.

1.1 Literature Review

In Table 1, we provide a brief comparison with the main works we have referred to. Each of these works proposed an algorithm for finding near-optimal policies for the discounted case.

One of the earliest results on approximating the optimal value function of a discounted MDP using the value function of a DMDP, defined by a sampled with generative model (simulator) transition probability matrix , was obtained in Gheshlaghi Azar et al. (2013). In this result, the authors accounted for the fact that the empirical Bellman equation might not be solved exactly. Specifically, they provided a bound on the norm of the difference between the optimal value function of the theoretical MDP and the value function obtained when solving the empirical Bellman equation using some iterative method at a certain step, with the bound depending on the iteration number. However, this bound was obtained only for two specific methods, namely value iteration and policy iteration, making it not directly applicable to other methods.

Meanwhile, in a more recent result Zurek & Chen (2023), a state-of-the-art bound was derived, where the authors managed to eliminate the prefactor $\frac{1}{(1-\gamma)^3}$ and replace it with $\frac{H}{(1-\gamma)^2}$, where $H \leq \frac{1}{1-\gamma}$. However, their result assumes knowledge of this parameter H, which may be unknown in practice. This limitation was addressed by the authors in Tuynman et al. (2024), where they attempted to account for this shortcoming. In a similar approach, the article Wang et al. (2023a) obtained a bound through an introduced characteristic of the environment called minorization time, and their result asymptotically required fewer samples. However, the bound required uniform ergodicity, which is not as general. Nonetheless, most of these works assume the ability to find an exact solution to the Bellman equation. Utilizing some ideas from the mentioned works, we aim to simultaneously address two factors: the inaccuracy of the empirical kernel relative to the theoretical one, which is $\|\hat{P} - P\|_\infty$ and the inaccuracy in solving the Bellman equation.

Reference	**Sample Complexity**	**Takes Into Account Inaccuracy in Solution**
Gheshlaghi Azar et al. (2013)	$\tilde{O}\left(\frac{\lvert S\rvert\lvert A\rvert}{(1-\gamma)^3\epsilon^2}\right)$	✓
Wang et al. (2023a)	$\tilde{O}\left(\frac{\lvert S\rvert\lvert A\rvert t_{\text{minorize}}}{(1-\gamma)^2\epsilon^2}\right)$	×
Zurek & Chen (2023)	$\tilde{O}\left(\frac{\lvert S\rvert\lvert A\rvert H}{(1-\gamma)^2\epsilon^2}\right)$	×
This paper	$\tilde{O}\left(\frac{\lvert S\rvert\lvert A\rvert H}{(1-\gamma)^2\epsilon^2}\right)$	✓

Table 1: Comparison of algorithms based on sample complexity.

1.2 PROBLEM SETUP

A Markov Decision Process is defined by a set (S, A, P, r), where S is a finite set of states, A is a finite set of actions, P: $S \times A \to \Delta(S)$ is the transition kernel, where $\Delta(S)$ denotes the probability simplex over the state space, and r: $S \times A \to [0; 1]$ is the reward function. In this work, we only deal with stationary strategies π: $S \to \Delta(A)$. We introduce the transition probability matrix induced by the policy π as P_π, defined by the formula:

$$(P_\pi)_{s,s'} = \sum_{a \in A} \pi(a|s) P(s'|s, a) \tag{1}$$

Similarly, the convolution of the policy with the reward function yields a reward induced by some policy:

$$(r_\pi)_{s,s'} = \sum_{a \in A} \pi(a|s) r(s, a) \tag{2}$$

We denote the expectation for a given policy starting from state s_0 as $\mathbb{E}^\pi_{s_0}$. If the subscript is not specified, the component-wise expectation is implied, applied to the random variable. The same is done for the variance - the variance for a specific starting state is denoted by $\mathbb{V}^\pi_s(X) = \mathbb{E}^\pi_s \left(X - \mathbb{E}^\pi_s(X)\right)^2$, and in vector form as $\mathbb{V}^\pi$: $(\mathbb{V}^\pi[X])_s = \mathbb{V}^\pi_s(X)$.
The discounted MDP is defined by the set (S, A, P, r, γ), where γ is the discount factor. Upon defining the DMDP, such an object as the value function for a certain policy $V^\pi : S \to \mathbb{R}$ is introduced, which by definition is:

$$V^\pi(s) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \right], \tag{3}$$

$r_t = r(s_t, a_t)$ - the reward obtained at step t. A policy π^* is called optimal if for all $s \in S$, $\pi \to V^{\pi^*} \geq V^\pi$. Extremely important for our analysis is the variance of the value function determined by the same policy, i.e., $\mathbb{V}^\pi[V^\pi] = P_\pi[V^\pi - P_\pi V^\pi]^2$, which component-wise can be written as:

$$(\mathbb{V}^\pi[V^\pi])_s = (P_\pi)_{s,s'} \left(V^\pi(s') - (P_\pi)_{s,s''} V^\pi(s'')\right)^2 \tag{4}$$

Finally, for AMDP, in which the reward is defined as the average of the mathematical expectation, the quantity $\rho^\pi(s) = \lim_{T \to \infty} \mathbb{E}^\pi \left[\frac{1}{T} \sum_{t=0}^{T-1} r_t \right]$ is introduced. Similarly, the Bellman/Poisson equation is introduced:

$$r_\pi - \rho^\pi = (I - P_\pi) h^\pi, \tag{5}$$

where $h^\pi(s) = \text{C-lim}_{T \to \infty} \mathbb{E}^\pi_s \left[\sum_{t=0}^{T-1} (r_\pi(s_t) - \alpha^\pi) \right]$ - the so-called bias function (here the Cesàro limit is implied). The solution of this equation is described by the pair (α^π, h^π), $\alpha^\pi \in \mathbb{R}$, $h^\pi : \mathbf{S} \to \mathbb{R}$. It is easy to see that this solution is not unique, and any pair from the set $\{(\alpha^\pi, h^\pi + ce) : c \in \mathbb{R}\}$, where $e(s) = 1 \; \forall s \in \mathbf{S}$, is also a solution. Analogously to the discounted case, the optimal policy $\pi^* = \arg\max_\pi \alpha^\pi$ is introduced. Recalling the basic concepts, we can now introduce another parameter $H := \|h^{\pi^*}\|_{span} := \max_s h^{\pi^*}(s) - \min_s h^{\pi^*}(s)$ - the span semi-norm of the optimal bias function. Note that, unlike the solution of the Poisson equation for a certain policy, this parameter is uniquely defined.

It is important to note that in existing results, sample complexity has been characterized using various parameters, such as the diameter of the MDP, the minorization time, the uniform mixing time bound τ_{unif}, and the previously mentioned span H of the optimal bias. This work relies heavily on the study in Zurek & Chen (2023), which analyzes sample complexity in terms of H.

The span H is particularly advantageous because it is always finite in weakly communicating MDPs with finite state-action spaces. Unlike the diameter D or the uniform mixing time τ_{unif}, which can each be arbitrarily larger than the other and can even be infinite under certain conditions, H remains bounded. For instance, it has been shown that $H \leq D$ Bartlett & Tewari (2012) and $H \leq 8\tau_{\text{unif}}$ Wang et al. (2022), which provides useful insights into the relative magnitudes of these parameters.

Furthermore, sample complexity bounds that involve τ_{unif} necessitate the assumption that all stationary policies have finite mixing times. If this assumption is not met, τ_{unif} becomes infinite, as in cases where any policy induces a periodic Markov chain. Similarly, while the diameter D can also be infinite, the span H is always finite, ensuring a more robust and reliable parameter for analysis.

2 MAIN RESULT

As previously mentioned, we are working with a generative model (simulator). Assuming access to this generative model, we collect n independent samples

$$s^i_{s,a} \overset{\text{i.i.d.}}{\sim} P(\cdot|s,a), \text{ for } i = 1, \ldots, n \tag{6}$$

for each pair $(s, a) \in S \times A$, which allows us to construct an empirical transition probability matrix (tensor)

$$\forall s' \in S, \hat{P}(s'|s,a) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{s^i_{s,a} = s'\}\,, \tag{7}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Thus, $\hat{P}(s'|s,a)$ calculates the empirical frequency of transitions from (s, a) to s'. Therefore, the total number of samples equals $n|S||A|$, and we can introduce an MDP defined by the set $(S, A, \hat{P}, r, \gamma)$. For this MDP, the value function $\hat{V}^\pi$ determined by some policy π is also defined.

The task of finding an ε-optimal policy for AMDP reduces to the search for a $\tilde{\varepsilon}$-optimal policy for DMDP. Our goal is to obtain an upper bound on $\|V^* - V^{\pi_t}\|_\infty$, which is likely to hold depending on the number of samples n. As previously noted, in order to find an ε-optimal policy for DMDP, we face the necessity of solving the empirical Bellman equation, and for the case when the discount factor is close to one, this is a separate problem. Various approaches have been considered for this case, such as in Goyal & Grand-Clement (2023); Grand-Clément (2021), drawing analogies with well-known iterative methods from convex optimization, as well as more heuristic methods, as in Farahmand & Ghavamzadeh (2021), which can provide improved results, but for which a theoretical result in the general case is currently lacking. Our result is inspired by an attempt to address this subproblem that arises when trying to find an approximately optimal policy for AMDP, for which we allow for some inaccuracy in the solution, depending on the iteration and the method itself, and account for it in our bound.

Before proceeding to the main result, we present the most fundamental lemmas used in deriving the bound, while the rest will be mentioned directly in the proof of our theorem. One of these lemmas relies on Bernstein's inequality, a concentration inequality from probability theory. It has been used in several results, but for completeness, its proof can also be found below (see Appendix A.1).

Lemma 1 (Bernstein's inequality)**.** *Let V be a value function that does not depend on the outcome of sampling. Then for any policy π, with probability at least $1 - \delta$, the following inequality holds:*

$$|(P_\pi - \hat{P}_\pi)V| \leq \sqrt{\frac{\beta}{n}\mathbb{V}^\pi[V]} + \frac{2\beta\|V\|_\infty}{3n}\mathbf{1}\,, \tag{8}$$

where $\beta = 2\log(\frac{2|S||A|}{\delta})$. Briefly, we will also mention two main lemmas that we relied on to obtain our result, which are well-known, so we will only reference them.

Lemma 2 ([Weissman et al. (2003), Lemma 13])**.** *Let p_z and $\hat{p}_z$ be probability distributions over a finite set of states $\mathbf{S}$. Then with probability at least $1 - \delta$, the following inequality holds:*

$$\|p_z - \hat{p}_z\|_1 \leq \sqrt{\frac{2|S|\log\left(\frac{2}{\delta}\right)}{n}} \tag{9}$$

Lemma 3 ([Singh & Yee (1994), Theorem])**.** *Let V^* be the optimal value function of a discounted MDP with discount factor γ, and V_t be a value function such that the inequality $\|V^* - V_t\|_\infty \leq \varepsilon$ holds. Let π_t be the greedy policy derived from V_t. Then the following bound holds:*

$$\|V^* - V^{\pi_t}\|_\infty \leq \frac{2\gamma\varepsilon}{1-\gamma} \tag{10}$$

Finally, we are ready to present our algorithm and guarantees for it.

Theorem 1. *The policy obtained by Algorithm 1 for the given DMDP is ε-optimal:*

$$\|V^* - V^{\pi_t}\|_\infty \leq \varepsilon + \frac{1}{(1-\gamma)\eta}\|\hat{V}^*_p - V_t\|_\infty, \tag{11}$$

Algorithm 1 Perturbed Model-Based Planning

Input: Parameter $\eta \in (0,1)$, sample size per state-action pair $n \geq \frac{500H}{(1-\gamma)^2\varepsilon^2\eta^4}\beta$, target accuracy $\varepsilon \in \left(0, \frac{1-\eta}{\frac{1}{5}+(2-\eta)\sqrt{\frac{|\mathbf{S}|}{500H}}}\right]$, discount factor γ

1: **for each** state-action pair $(s,a) \in S \times A$ **do**
2: Collect n samples $s_1^{s,a}, \ldots, s_n^{s,a}$ from $P(\cdot|s,a)$
3: Form the empirical transition kernel $\hat{P}(s'|s,a) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{s_i^{s,a} = s'\}$, for all $s' \in S$
4: **end for**
5: Set perturbation level $\xi = \frac{(1-\gamma)\varepsilon\eta}{4}$
6: Form perturbed reward $\tilde{r} = r + Z$ where $Z(s,a) \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,\xi)$
7: Initialize the starting point(s) in $\mathbb{R}^{|S|}$ for running the iterative method
8: Compute a greedy policy π_T from the point V_T obtained at the T-th step
9: **return** π_T

Here are a few points to note: First, our result imposes a limitation on the required precision. As the number of states in the environment increases, the theoretical estimate remains valid for a smaller set of values of ε. This is similar to the limitation that arose in Gheshlaghi Azar et al. (2013). Second, it is interesting to note that in our derivation, we did not have to resort to the method of statistical independence, which has been actively applied in recent results. Finally, it might seem that this derivation introduces another hyperparameter of unclear nature, η. However, it is simply a certain number used to estimate the denominator value (see Appendix A.2). It can be chosen based on the desired precision in the search for a suboptimal policy.

3 Experiment

3.1 Setup

To begin with, let's clarify the environment we are working with. It is a square grid of size 21×21. Every cell can be represented as a tuple (x, y), where $x, y \in \overline{0, 20}$, with the first number corresponding to the row and the second to the column. Thus, the terminal state is $(20, 20)$. Also, there 4 actions are available: move down, up, right, and left, respectively (see Figure 1). However, this environment is not deterministic – when choosing an action, say, to move right, there is a probability of 0.2 of ending up in the adjacent upper cell or the same probability of ending up in the adjacent lower cell. Finally, the reward function is this:

$$r(s,a) = \begin{cases} 1 & \text{if } s = (20,19) \text{ and } a = \rightarrow \\ 1 & \text{if } s = (19,20) \text{ and } a = \downarrow \\ 0 & \text{otherwise} \end{cases}$$

The reason why it is advisable to consider this example lies in the large number of states and the presence of stochasticity, which may require many samples for each state-action pair to accurately approximate the environment with an empirical kernel.

In this experiment, we aim to observe how the error, i.e., the norm of the difference between the optimal value function of our MDP and the value function induced by some policy obtained from the empirical Bellman equation, behaves depending on the number of samples n. For this purpose, we developed a custom module for solving the Markov Decision Process. It includes methods for creating a random tabular environment, solving the optimality equation using various iterative methods (some of which are not included in the experiment), and sampling to obtain the empirical kernel.

It was not mentioned earlier, but it should be understood that even in the case where there is no discrepancy between the theoretical and empirical transition kernels, some error will persist. This is due to the need to add some perturbation to the reward $\xi \sim \text{Uniform}(0, \zeta)$ to obtain a unique policy Li et al. (2020). We note that, firstly, the amplitude of the perturbation is scaled by the factor $1 - \gamma$, and secondly, by the error ε that we aim to achieve.

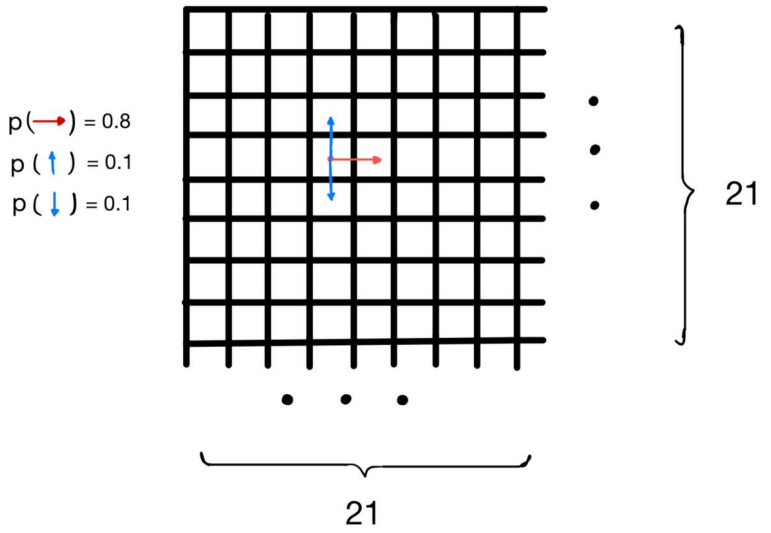


Figure 1: The environment

3.2 SEARCH OF H

In the theoretical estimate, the parameter H is present, and it can be found in several ways. For this parameter, as well as for other parameters such as t_{mix}, t_{minorize}, etc., which were used in previous works, theoretical estimates have been obtained. However, to verify our result, we would like to find the exact value of this parameter, and this can be done.

Firstly, we note that the following is only valid for the case of uniform ergodicity. This problem can be formulated as a linear programming task - we want to find a policy that maximizes

$$\rho^{\pi} = \lim_{H\to\infty} \frac{1}{H} \mathbb{E}^{\pi} \left[\sum_{t=0}^{H-1} r(s_t, a_t) \right] = \sum_{(s,a)\in S\times A} r(s,a)\pi(a|s)\nu_{\pi}(s) \tag{12}$$

The last inequality holds due to uniform ergodicity. Here, $\nu_{\pi}(s)$ is the stationary distribution of the Markov process, which can be expressed as

$$\nu_{\pi}(s') = \sum_{(s,a)\in S\times A} p(s|s',a)\pi(a|s)\nu_{\pi}(s)\,, \tag{13}$$

and corresponds to its probability vector $\nu_{\pi} = (\nu_{\pi}(s))_{s\in S}$. From the theory of stochastic processes, it is known that the probability vector corresponding to the stationary distribution can be determined as the eigenvector of the transposed stochastic matrix that defines the Markov chain. Given the formula for P_{π}, the formula for $\nu_{\pi}(s')$ becomes evident.
Returning to the search for the optimal policy, we introduce the distribution of actions over states $\mu(s,a) = \nu_{\pi}(s)\pi(a|s)$, then we can rewrite the policy search problem in AMDP as an LP problem with the sense of policy value estimation by distribution μ:

$$\max_{\mu\in\Delta^{\mathbf{S}\times\mathbf{A}}} \left[\rho(\mu) = \sum_{(s,a)\in\mathbf{S}\times\mathbf{A}} r(s,a)\mu(s,a) : \sum_{b\in\mathbf{A}} \mu(s',b) = \sum_{(s,a)\in\mathbf{S}\times\mathbf{A}} p(s'|s,a)\mu(s,a), s'\in\mathbf{S} \right], \tag{14}$$

where $\Delta^{\mathbf{S}\times\mathbf{A}} = \{\mu : \mu(s,a) \geq 0, \sum_{(s,a)\in\mathbf{S}\times\mathbf{A}} \mu(s,a) = 1\}$.
Knowing the distribution μ, we can recover the optimal policy $\pi_{\mu}(a|s) = \frac{\mu(s,a)}{\sum_{b\in\mathbf{A}}\mu(s,b)}$. This problem can be rewritten in matrix form:

$$\max_{\mu\in\Delta^{\mathbf{S}\times\mathbf{A}}} \langle r, \mu\rangle \tag{15}$$

$$\text{s.t. } (\hat{I} - P)\mu = 0 \tag{16}$$

The identity matrix $\hat{I}$ has a non-standard format: it is a rectangular matrix of size $|S| \times |S||A|$, with only the elements corresponding to the pair (s, a), $a \in A$, equal to one on each row $s \in S$, while the other elements of this row are zero, i.e., there are exactly $|A|$ ones in each row of $\hat{I}$. The matrix P of size $|S| \times |S||A|$ has the distribution $P(\cdot|s, a)$ in each column $(s, a) \in S \times A$.
In general, after this, it would be necessary to find the stationary distribution of the Markov process, determine ρ^π, and substitute this value into the Poisson equation to finally find the bias function and, accordingly, its span semi-norm. However, we would like to directly find this bias function. For this LP problem, the dual problem is directly constructed, under the condition that $\mu \geq 0$, which makes sense in evaluating the value of the optimal policy through the h-function:

$$\min_{\rho \in \mathbb{R}, h \in \mathbb{R}^{\mathbf{S}}} \rho \tag{17}$$

$$\text{s.t. } r - \rho \mathbf{1} - (\hat{I} - P)h \leq 0 \tag{18}$$

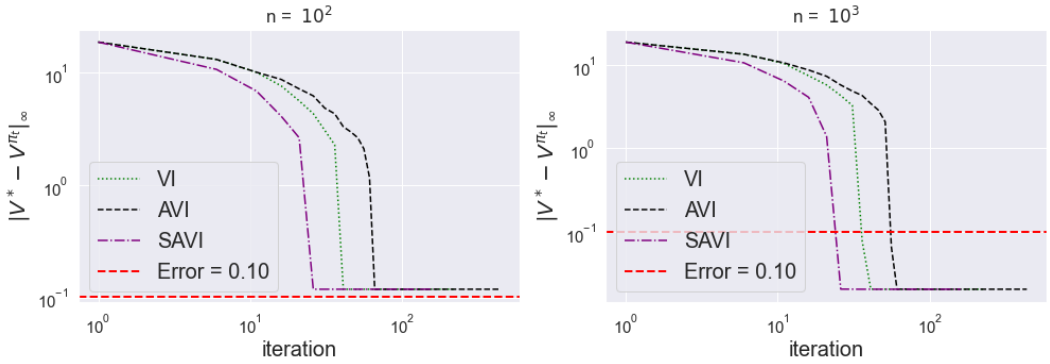
Thus, the Bellman optimality equation

$$h(s) = \max_{a \in A} [r(s, a) - \rho^* + \sum_{s' \in S} p(s'|s, a)h(s')] \,, \tag{19}$$

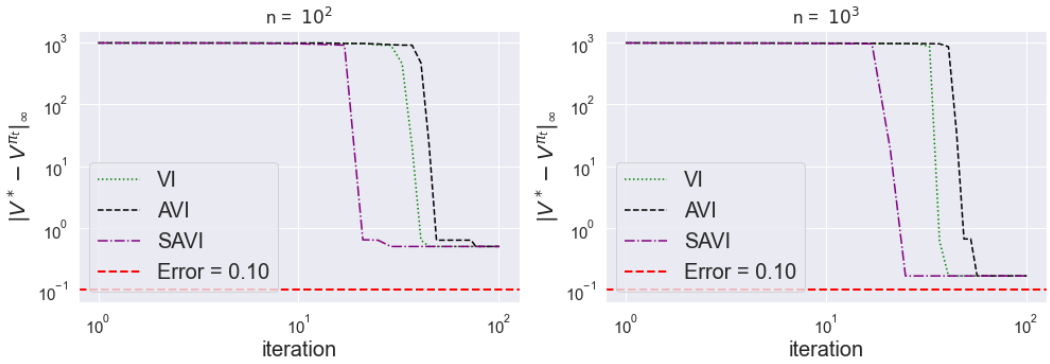
is obtained from the constraints of the form $\hat{I}^T h \geq r - \rho \mathbf{1} + P^T h$. We will use this approach as it helps directly find both the reward induced by the optimal policy and the parameter H.

3.3 Results

Finally, we can proceed to the discussion of the results (see Figure 2). First and foremost, it is worth noting that due to the factor $\frac{1}{1-\gamma}$, the initial error differs by an order of magnitude for $\gamma = 0.95$ and $\gamma = 0.999$. The same applies to the plateau reached by the residual. For example, in the first case, only $n = 10^2$ samples are sufficient for each pair $(s, a) \in S \times A$, while in the second case, $n = 10^3$ is still insufficient to achieve the desired error.



(a) $\gamma = 0.95$



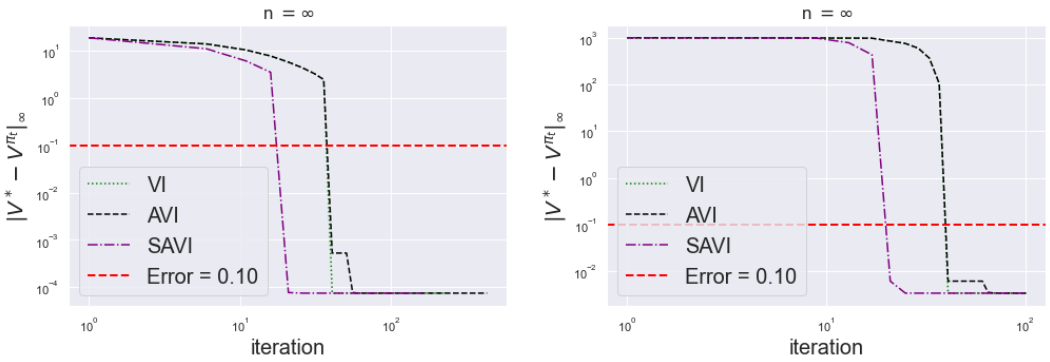
(b) $\gamma = 0.999$

Figure 2: Experiment set for different number of samples

An interesting point here is the convergence behavior towards the optimum (which we will discuss further in the next section). For the case where the discount factor is close to one, the convergence plot looks like a staircase. Although our environment requires relatively few iterations to converge to the solution, for some so-called hard cases, this can be a problem, as an insufficient number of steps may not effectively reduce the initial norm of the value function differences.

For comparison, let's see what inaccuracy is achieved in the limiting case when the number of samples for each state-action pair tends to infinity (see Figure 3). For this, instead of the empirical kernel, we simply substitute the theoretical one. The graphs obtained for intermediate values of the number of samples can be found in Appendix B.

We also assumed in the proof of Algorithm 1 that $\gamma \geq 1 - \frac{1}{H}$. In our setup, H reached a value of 5.12. However, there is no explicit limitation on H, so the value of this hyperparameter can reach values such that $\frac{1}{H} \ll 1$, which means the obtained result will only be valid for DMDPs with a discount factor close to one. Since the Bellman operator is a contraction with the coefficient γ, the estimate $\|V^* - V_t\|_\infty \leq \gamma^t \|V^* - V_0\|_\infty$ holds, which means that to ensure the required accuracy, it may potentially take $T \sim \frac{1}{1-\gamma}$ iterations of Value Iteration (VI). Thus, convergence to the solution of the Bellman equation using the standard iterative method becomes slow, so even with a large number of iterations, the residual in the solution can be significant, and it should be considered in the estimate of $\|V^* - V^{\pi_t}\|_\infty$.



(a) $\gamma = 0.95$ (b) $\gamma = 0.999$

Figure 3: Case of infinite number of samples

Therefore, it seems reasonable to apply other methods for finding the optimal value function. The main ones are mentioned in Grand-Clément (2021). For instance, for Accelerated Value Iteration (AVI), it is shown that the iterative complexity of the algorithm is $T \sim \frac{1}{\sqrt{1-\gamma}}$. Moreover, one of them, Safe-Accelerated Value Iteration (S-AVI), has consistently shown improved results. However, to some extent, it can be considered heuristic, as there is no estimate proving that it is more efficient than conventional VI.

4 Conclusion

In this work, we have advanced the study of ε-optimal policies for Average Reward Markov Decision Processes (AMDP) by reducing them to Discounted Markov Decision Processes (DMDP). Unlike existing research, which often requires the discount factor to be close to one, our approach allows for inaccuracies in solving the empirical Bellman equation. Despite this allowance, we have maintained sample complexity that aligns with the latest results. By separating the contributions from the inaccuracy of approximating the transition matrix and the residuals in solving the Bellman equation, we have enabled a more accurate determination of the total complexity of ε-optimal policy analysis for DMDP across various iterative methods with theoretical foundations in iterative complexity.

Additionally, we conducted experiments using different iterative methods for solving the Bellman equation, revealing that the error can remain nearly constant relative to the initial approximation for

several steps. This insight is crucial, particularly when a large number of steps are required, as it highlights the importance of accounting for residuals in solving the equation. In future work, we aim to generalize our estimates further by expanding the range of permissible values for desired accuracy in finding near-optimal policies.

REFERENCES

Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661*, 2012.

Amir-massoud Farahmand and Mohammad Ghavamzadeh. Pid accelerated value iteration algorithm. In *International Conference on Machine Learning*, pp. 3143–3153. PMLR, 2021.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91: 325–349, 2013.

Vineet Goyal and Julien Grand-Clement. A first-order approach to accelerated value iteration. *Operations Research*, 71(2):517–535, 2023.

Julien Grand-Clément. From convex optimization to mdps: A review of first-order, second-order and quasi-newton methods for mdps. *arXiv preprint arXiv:2104.10677*, 2021.

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

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in neural information processing systems*, 33:12861–12872, 2020.

Tianjiao Li, Feiyang Wu, and Guanghui Lan. Stochastic first-order methods for average-reward markov decision processes. *arXiv preprint arXiv:2205.05800*, 2022.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Satinder P Singh and Richard C Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16:227–233, 1994.

Daniil Tiapkin and Alexander Gasnikov. Primal-dual stochastic mirror descent for mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 9723–9740. PMLR, 2022.

Adrienne Tuynman, Rémy Degenne, and Emilie Kaufmann. Finding good policies in average-reward markov decision processes without prior knowledge. *arXiv preprint arXiv:2405.17108*, 2024.

Jinghan Wang, Mengdi Wang, and Lin F Yang. Near sample-optimal reduction-based policy learning for average reward mdp. *arXiv preprint arXiv:2212.00603*, 2022.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward markov decision processes. *arXiv preprint arXiv:2310.08833*, 2023a.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity of reinforcement learning for mixing discounted markov decision processes. *arXiv preprint arXiv:2302.07477*, 2023b.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, pp. 125, 2003.

Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for average reward mdps. *arXiv preprint arXiv:2311.13469*, 2023.

A PROOFS

A.1 PROOF OF LEMMA 1

Proof. First, we estimate the quantity $\mathbb{P}(|(\hat{p}_z - p_z) \cdot V| > \varepsilon), V \in \mathbb{R}^{|\mathbf{S}|}, z \in (\mathbf{S} \times \mathbf{A})$

$$\mathbb{P}\left((\hat{p}_z - p_z) \cdot V > \varepsilon\right) = \mathbb{P}\left(\sum_{i=1}^{n}\left[\sum_{s'}(X_i^{s'} - p_z^{s'})V^{s'}\right] > n\varepsilon\right) \tag{20}$$

where $X_i^{s'}$ denotes a Bernoulli indicator random variable that takes the value 1 if the i-th sample for the state-action pair z transitions to state s', and 0 otherwise.

$$\sigma_i^2 = \mathbb{E}[\sum_{s'}(X_i^{s'} - p_z^{s'})V^{s'}]^2 = \mathbb{E}[\sum_{s'}(X_i^{s'} - p_z^{s'})^2(V^{s'})^2] + \mathbb{E}[\sum_{s' \neq s''}(X_i^{s'} - p_z^{s'})(X_i^{s''} - p_z^{s''})V^{s'}V^{s''}] \tag{21}$$

$$\mathbb{E}[\sum_{s'}(X_i^{s'} - p_z^{s'})^2(V^{s'})^2] = \sum_{s'}(V^{s'})^2 p_z^{s'}(1 - p_z^{s'}) \tag{22}$$

$$\mathbb{E}[\sum_{s' \neq s''}(X_i^{s'} - p_z^{s'})(X_i^{s''} - p_z^{s''})V^{s'}V^{s''}] = -\sum_{s' \neq s''} V^{s'}V^{s''}p_z^{s'}p_z^{s''} \tag{23}$$

$$\sigma^2\left(\sum_{i=1}^{n}\left[\sum_{s'}(X_i^{s'} - p_z^{s'})V^{s'}\right]\right) = n\sigma_i^2 = n\left(\sum_{s'}(V^{s'})^2 p_z^{s'} - \left(\sum_{s'} V^{s'} p_z^{s'}\right)^2\right) := n\sigma^2(V)(z), \tag{24}$$

where σ_i^2 denotes the variance of a single term.

Applying Bernstein's inequality and bounding $|(p_z - \hat{p}_z)V| \leq 2\|V\|_\infty$, we obtain:

$$\mathbb{P}\left(|(\hat{p}_z - p_z) \cdot V| \leq \varepsilon\right) \geq 1 - 2\exp\left(-\frac{n\varepsilon^2}{2\sigma^2(V)(z) + \frac{4}{3}\varepsilon\|V\|_\infty}\right) \geq 1 - \delta$$

$$\Rightarrow \varepsilon \geq \frac{2\|V\|_\infty}{3n}\ln\frac{2}{\delta} + \sqrt{\left(\frac{2\|V\|_\infty}{3n}\ln\frac{2}{\delta}\right)^2 + \frac{2\sigma^2(V)(z)}{n}\ln\frac{2}{\delta}} \tag{25}$$

It is sufficient to take $\varepsilon \geq \frac{4\|V\|_\infty}{3n}\ln\frac{2}{\delta} + \sigma(V)(z)\sqrt{\frac{2\ln\frac{2}{\delta}}{n}}$. Now, everywhere in our reasoning, replace δ with $\frac{\delta}{|S||A|}$, use the union bound, and recalling the expression for $\mathbb{V}^\pi[V]$, we obtain the required assertion.

□

A.2 PROOF OF THEOREM 1

Proof. First, we need to group the terms so that each of them contains objects of the same nature.

$$V^* - V^{\pi_t} = V^* - \hat{V}_p^{\pi^*} + \hat{V}_p^{\pi^*} - \hat{V}_p^* + \hat{V}_p^* - \hat{V}_p^{\pi_t} + \hat{V}_p^{\pi_t} - V^{\pi_t} \leq$$

$$\leq \|V^* - \hat{V}^{\pi^*}\|_\infty + \|\hat{V}_p^* - \hat{V}_p^{\pi_t}\|_\infty + \|\hat{V}^{\pi_t} - V^{\pi_t}\|_\infty + \frac{\zeta}{1-\gamma} \tag{26}$$

Thus, the value functions in the first and third terms correspond to the same policy, and in the second term, they correspond to the same transition kernel.

Let us handle each of the three terms separately, starting with the most challenging one, namely $\|\hat{V}^{\pi_t} - V^{\pi_t}\|_\infty$. This term is of the greatest interest because our goal is to try to reduce the degree of the prefactor $\frac{1}{1-\gamma}$, which would yield an improved result. Therefore, the approach to evaluating

this term will differ from that in Gheshlaghi Azar et al. (2013). But first, let's perform the basic transformation:

$$\hat{V}^{\pi_t} - V^{\pi_t} = (I - \gamma P_{\pi_t})^{-1}(I - \gamma P_{\pi_t})\hat{V}^{\pi_t} - (I - \gamma P_{\pi_t})^{-1} r_{\pi_t} =$$
$$= (I - \gamma P_{\pi_t})^{-1}(I - \gamma P_{\pi_t})\hat{V}^{\pi_t} - (I - \gamma P_{\pi_t})^{-1}(I - \gamma \hat{P}_{\pi_t})\hat{V}^{\pi_t} =$$
$$= (I - \gamma P_{\pi_t})^{-1}(\hat{P}_{\pi_t} - P_{\pi_t})V^* \tag{27}$$

Let's represent $\hat{V}^{\pi_t}$ as $V^* + (\hat{V}^{\pi_t} - V^*)$, and then perform trivial inequalities:

$$\|\hat{V}^{\pi_t} - V^{\pi_t}\|_\infty \overset{(28.1)}{\leq} \gamma \|(I - \gamma P_{\pi_t})^{-1}(P_{\pi_t} - \hat{P}_{\pi_t})V^*\|_\infty + \frac{\gamma}{1-\gamma}\|(P_{\pi_t} - \hat{P}_{\pi_t})(V^* - \hat{V}^{\pi_t})\|_\infty \overset{(28.2)}{\leq}$$
$$\leq \gamma \|(I - \gamma P_{\pi_t})^{-1}(P_{\pi_t} - \hat{P}_{\pi_t})V^*\|_\infty + \frac{\gamma}{1-\gamma}\|P_{\pi_t} - \hat{P}_{\pi_t}\|_\infty \|V^* - \hat{V}^{\pi_t}\|_\infty \overset{(28.3)}{\leq}$$
$$\leq \gamma \|(I - \gamma P_{\pi_t})^{-1}(P_{\pi_t} - \hat{P}_{\pi_t})V^*\|_\infty + \frac{\gamma}{1-\gamma}\|P_{\pi_t} - \hat{P}_{\pi_t}\|_\infty (\|V^* - V^{\pi_t}\|_\infty + \|V^{\pi_t} - \hat{V}^{\pi_t}\|_\infty)\,, \tag{28}$$

where in (28.1) and (28.3) the triangle inequality was applied, and in (28.2) the property of the norm was used. Now, moving all terms with $\hat{V}^{\pi_t} - V^{\pi_t}$ to the right side, we obtain:

$$\|\hat{V}^{\pi_t} - V^{\pi_t}\|_\infty \leq \frac{\gamma}{1 - \frac{\gamma}{1-\gamma}\|P_{\pi_t} - \hat{P}_{\pi_t}\|_\infty} \Big(\|(I - \gamma P_{\pi_t})^{-1}(P_{\pi_t} - \hat{P}_{\pi_t})V^*\|_\infty +$$
$$+ \frac{1}{1-\gamma}\|P_{\pi_t} - \hat{P}_{\pi_t}\|_\infty \|V^* - V^{\pi_t}\|_\infty \Big) \tag{29}$$

Using Lemma 1 for the first term, we then obtain that with probability at least $1 - \delta$, the following inequality holds:

$$|(P_{\pi_t} - \hat{P}_{\pi_t})V^*| \leq \sqrt{\frac{\beta_1}{n}}\sqrt{\mathbb{V}^{\pi_t}[V^*]} + \frac{2\beta_1 \|V^*\|_\infty}{3n}\mathbf{1}\,, \tag{30}$$

where $\beta_1 = 2\log(\frac{|S||A|}{\delta})$. Then, using the property that $\forall \pi \to \|(I - \gamma P_\pi)^{-1}\|_\infty \leq \frac{1}{1-\gamma}$, we immediately arrive at the inequality:

$$\|(I - \gamma P_{\pi_t})^{-1}|(P_{\pi_t} - \hat{P}_{\pi_t})V^*|\|_\infty \leq \sqrt{\frac{\beta_1}{n}}\|(I - \gamma P_{\pi_t})^{-1}\sqrt{\mathbb{V}^{\pi_t}[V^*]}\|_\infty + \frac{2\beta_1}{3n(1-\gamma)^2} \tag{31}$$

It is also necessary to use the variance property

$$\sqrt{\mathbb{V}^{\pi_t}[V^*]} \leq \sqrt{\mathbb{V}^{\pi_t}[V^* - V^{\pi_t}]} + \sqrt{\mathbb{V}^{\pi_t}[V^{\pi_t}]} \leq \|V^* - V^{\pi_t}\|_\infty + \sqrt{\mathbb{V}^{\pi_t}[V^{\pi_t}]} \tag{32}$$

(that the square root of the variance of a sum is less than or equal to the sum of the square roots of the variances), which will help us to progress towards the estimate for
$\|(I - \gamma P_{\pi_t})^{-1}|(P_{\pi_t} - \hat{P}_{\pi_t})V^*|\|_\infty$:

$$\|(I - \gamma P_{\pi_t})^{-1}|(P_{\pi_t} - \hat{P}_{\pi_t})V^*|\|_\infty \leq \sqrt{\frac{\beta_1}{n}}\|(I - \gamma P_{\pi_t})^{-1}\sqrt{\mathbb{V}^{\pi_t}[V^{\pi_t}]}\|_\infty +$$
$$+ \sqrt{\frac{\beta_1}{(1-\gamma)^2 n}}\|V^* - V^{\pi_t}\|_\infty + \frac{2\beta_1}{3n(1-\gamma)^2} \tag{33}$$

The inequality for the variance was used to arrive at the known construction
$\|(I - \gamma P_\pi)^{-1}\sqrt{\mathbb{V}^\pi[V^\pi]}\|_\infty$. Now we can consistently apply [Zurek & Chen (2023), Lemma 6], [Zurek & Chen (2023), Lemma 7], [Zurek & Chen (2023), Lemma 8]:

$$\|(I - \gamma P_{\pi_t})^{-1}\sqrt{\mathbb{V}^{\pi_t}[V^{\pi_t}]}\|_\infty \leq \frac{1}{\gamma}\sqrt{\frac{2}{(1-\gamma)}}\sqrt{\left\|\mathbb{V}^{\pi_t}[\sum_{t=0}^{\infty}\gamma^t R_t]\right\|_\infty} \leq$$

$$\leq \frac{1}{\gamma}\sqrt{\frac{2}{(1-\gamma)}}\sqrt{\frac{\left\|\mathbb{V}^{\pi_t}[\sum_{t=0}^{H-1}\gamma^t R_t + \gamma^H V^{\pi_t}(S_H)]\right\|_\infty}{1-\gamma^{2H}}} \leq$$

$$\leq \frac{1}{\gamma}\sqrt{\frac{5}{2H(1-\gamma)^2}}\sqrt{\left\|\mathbb{V}^{\pi_t}[\sum_{t=0}^{H-1}\gamma^t R_t + \gamma^H V^{\pi_t}(S_H)]\right\|_\infty} \tag{34}$$

Since π_t is not the optimal policy here, [Zurek & Chen (2023), Lemma 9] cannot be directly applied, but we can do something similar:

$$\left\|\mathbb{V}^{\pi_t}\left[\sum_{t=0}^{H-1}\gamma^t R_t + \gamma^H V^{\pi_t}(S_H)\right]\right\|_\infty \leq 3\mathbb{E}^{\pi_t}\left|\sum_{t=0}^{H-1}\gamma^t R_t\right|^2 + 3\mathbb{E}^{\pi_t}\left|\gamma^H(V^{\pi_t} - V^*)\right|^2 +$$

$$+ 3\mathbb{E}^{\pi_t}\left|\gamma^H\left(V^* - \frac{1}{1-\gamma}\rho^*\right)\right|^2 \leq 6H^2 + 3\|V^* - V^{\pi_t}\|_\infty^2. \tag{35}$$

In this derivation, we used two properties. The first is that the variance $\mathbb{V}(X) \leq \mathbb{E}(X^2)$ for some random variable, and the second is that $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$.

Finally, using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we obtain an estimate for $\|(I-\gamma P_\pi)^{-1}\sqrt{\mathbb{V}^\pi[V^\pi]}\|_\infty$:

$$\|(I-\gamma P_{\pi_t})^{-1}\sqrt{\mathbb{V}^{\pi_t}[V^{\pi_t}]}\|_\infty \leq \frac{1}{\gamma}\sqrt{\frac{15H}{(1-\gamma)^2}} + \frac{1}{\gamma}\sqrt{\frac{15}{2H(1-\gamma)^2}}\|V^* - V^{\pi_t}\|_\infty \tag{36}$$

Denoting $\frac{\gamma}{1-\gamma}\|P_{\pi_t} - \hat{P}_{\pi_t}\|_\infty$ by x, we obtain the final estimate for $\|\hat{V}^{\pi_t} - V^{\pi_t}\|_\infty$:

$$\|\hat{V}^{\pi_t} - V^{\pi_t}\|_\infty \leq \frac{1}{1-x}\left[\gamma\left(\sqrt{\frac{\beta_1}{n}}\|(I-\gamma P_{\pi_t})^{-1}\sqrt{\mathbb{V}^{\pi_t}[V^{\pi_t}]}\|_\infty + \sqrt{\frac{\beta_1}{(1-\gamma)^2 n}}\|V^* - V^{\pi_t}\|_\infty + \right.\right.$$

$$\left.\left. + \frac{2\beta_1}{3n(1-\gamma)^2}\right) + x\|V^* - V^{\pi_t}\|_\infty\right] \leq \frac{1}{1-x}\left(\sqrt{\frac{15H\beta_1}{(1-\gamma)^2 n}} + \sqrt{\frac{15\beta_1}{2Hn(1-\gamma)^2}}\|V^* - V^{\pi_t}\|_\infty + \right.$$

$$\left. + \sqrt{\frac{\beta_1}{(1-\gamma)^2 n}}\|V^* - V^{\pi_t}\|_\infty + \frac{2\beta_1}{3n(1-\gamma)^2} + x\|V^* - V^{\pi_t}\|_\infty\right) \tag{37}$$

Using the fact that $\beta_1 \leq \beta = 2\log(\frac{2|S||A|\log(\frac{e}{1-\gamma})}{\delta})$, we obtain the final estimate for $\|V^* - \hat{V}^{\pi_t}\|_\infty$:

$$\|V^* - V^{\pi_t}\|_\infty \leq \frac{\|V^* - \hat{V}^{\pi^*}\|_\infty + \|\hat{V}_p^* - V_t\|_\infty + \frac{\zeta}{1-\gamma} + \frac{1}{1-x}\left[\frac{2\beta}{3n(1-\gamma)^2} + \sqrt{\frac{15H\beta}{(1-\gamma)^2 n}}\right]}{1 - \frac{1}{1-x}\left(x + \sqrt{\frac{15\beta}{2Hn(1-\gamma)^2}} + \sqrt{\frac{\beta}{(1-\gamma)^2 n}}\right)} \tag{38}$$

Let us explain how we moved from $\|\hat{V}_p^* - \hat{V}_p^{\pi_t}\|_\infty$ to $\|\hat{V}_p^* - V_t\|_\infty$. We applied Lemma 3, taking $\varepsilon = \|\hat{V}_p^* - V_t\|_\infty$, so that the inequality in the condition holds trivially. Thus, we obtain:

$$\|\hat{V}_p^* - \hat{V}_p^{\pi_t}\|_\infty \leq \frac{2\gamma}{1-\gamma}\|\hat{V}_p^* - V_t\|_\infty \leq \frac{2}{1-\gamma}\|\hat{V}_p^* - V_t\|_\infty \tag{39}$$

Let's handle the term $\|V^* - V^{\pi^*}\|_\infty$ - since π^* does not depend on the sampling result, we can directly use the result from [Li et al. (2020), Lemma 1]:

$$\|V^* - V^{\pi^*}\|_\infty \leq 4\gamma\sqrt{\frac{\beta}{n}}\left\|(I-\gamma P_{\pi^*})^{-1}\sqrt{\mathbb{V}^{\pi^*}[V^*]}\right\|_\infty + \gamma\frac{\beta}{(1-\gamma)n}\|V^*\|_\infty\,, \tag{40}$$

where $n \geq \frac{16e^2}{1-\gamma}\beta$. We then perform the same steps as in the previous estimate and obtain:

$$\|V^* - V^{\pi^*}\|_\infty + \frac{\zeta}{1-\gamma} \leq 4\gamma\sqrt{\frac{\beta}{n}}\sqrt{\frac{10}{(1-\gamma)^2}H} + \gamma\frac{\beta}{(1-\gamma)^2 n} \tag{41}$$

Let $n \geq \frac{500H}{(1-\gamma)^2\varepsilon^2}\beta$, then we get the following estimate:

$$\|V^* - V^{\pi_t}\|_\infty \leq \frac{\left(\sqrt{\frac{160}{500}} + \frac{1}{500}\right)\varepsilon + \frac{1}{1-\gamma}\|\hat{V}_p^* - V_t\|_\infty + \frac{\zeta}{1-\gamma} + \frac{\varepsilon}{1-x}\left(\frac{2}{1500} + \sqrt{\frac{15}{500}}\right)}{1 - \frac{1}{1-x}\left(x + \frac{\varepsilon}{5}\right)} \tag{42}$$

Now apply Lemma 2, which gives us:

$$\|P_{\pi_t} - \hat{P}_{\pi_t}\|_\infty \leq \sqrt{\frac{|S|\beta}{n}} \Rightarrow x \leq \sqrt{\frac{|S|\beta}{(1-\gamma)^2 n}} \leq \varepsilon\sqrt{\frac{|S|}{500H}} := \mathcal{C}\varepsilon \tag{43}$$

where we used the condition on n. Recall that at least 2 conditions apply to x:

$$\left.\begin{aligned} 1 - x \geq 1 - \mathcal{C}\varepsilon \geq \eta \\ 1 - \frac{\frac{\varepsilon}{5} + x}{1-x} \geq \eta \end{aligned}\right\} \implies \left.\begin{aligned} \varepsilon \leq \frac{1-\eta}{\mathcal{C}} \\ \varepsilon \leq \frac{1-\eta}{\frac{1}{5} + (2-\eta)\mathcal{C}} \end{aligned}\right\} \implies \varepsilon \leq \frac{1-\eta}{\frac{1}{5} + (2-\eta)\mathcal{C}}, \tag{44}$$

where $\eta \in (0; 1)$. In total, the upper bound for $\|V^* - V^{\pi_t}\|_\infty$ takes the form:

$$\|V^* - V^{\pi_t}\|_\infty \leq \frac{\left(\sqrt{\frac{160}{500}} + \frac{1}{500}\right)\varepsilon + \frac{1}{1-\gamma}\|\hat{V}_p^* - V_t\|_\infty + \frac{\zeta}{1-\gamma} + \frac{\varepsilon}{\eta}\left(\frac{2}{1500} + \sqrt{\frac{15}{500}}\right)}{\eta} \tag{45}$$

Let $n \geq \frac{500H}{(1-\gamma)^2\varepsilon^2\eta^4}\beta$:

$$\|V^* - V^{\pi_t}\|_\infty \leq \frac{3\varepsilon}{4} + \frac{\zeta}{(1-\gamma)\eta} + \frac{1}{(1-\gamma)\eta}\|\hat{V}_p^* - V_t\|_\infty \tag{46}$$

Finally, we can take $\zeta = \frac{(1-\gamma)\eta\varepsilon}{4}$, which gives:

$$\forall\eta \in (0;1)\ \forall\varepsilon \in \left(0; \frac{1-\eta}{\frac{1}{5} + (2-\eta)\sqrt{\frac{|\mathbf{S}|}{500H}}}\right) : \|V^* - V^{\pi_t}\|_\infty \leq \varepsilon + \frac{1}{(1-\gamma)\eta}\|\hat{V}_p^* - V_t\|_\infty$$

provided $n \geq \frac{500H}{(1-\gamma)^2\varepsilon^2\eta^4}\beta$ □

B EXPERIMENT

We examined the convergence to a solution using three iterative methods: the standard value iteration, the accelerated value iteration, and the momentum-based method.

For clarity, we will demonstrate the functionality from Algorithm 1 for each method and specify the hyperparameters used for them.

Algorithm 2 Accelerated Value Iteration

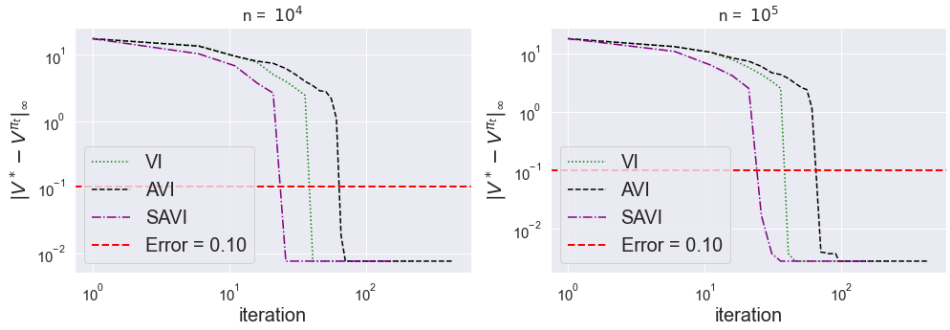
input: $\alpha = \frac{1}{1+\gamma}, \beta = \frac{1-\sqrt{1-\gamma^2}}{\gamma}$
Initialize $v_0 = \mathbf{0}$, $v_1 = T(v_0)$
for $t \in \{1, 2, ..., T-1\}$ **do**
    $h_t = v_t + \beta \cdot (v_t - v_{t-1})$
    $v_{t+1} = h_t - \alpha(h_t - T(h_t))$
end for

Algorithm 3 Safe Accelerated Value Iteration

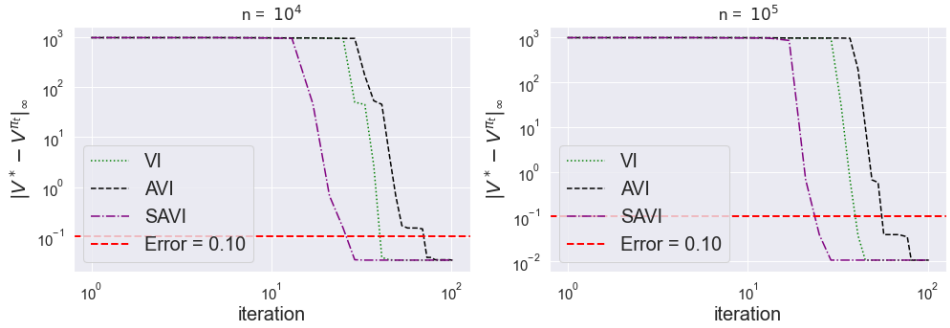
input: $\alpha = \frac{2}{1+\sqrt{1-\gamma^2}}, \beta = (\frac{\sqrt{1+\gamma}-\sqrt{1-\gamma}}{\sqrt{1+\gamma}+\sqrt{1-\gamma}})^2, \lambda = \frac{1+\gamma}{2}$
Initialize $v_0 = \mathbf{0}$, $v_1 = T(v_0)$
for $t \in \{1, 2, ..., T-1\}$ **do**
    Set
$$\begin{cases} h_t = v_t + \gamma \cdot (v_t - v_{t-1}) \\ v_{t+1/2} \leftarrow h_t - \alpha(h_t - T(h_t)) \end{cases}$$
    if $\|v_{t+1/2} - T(v_{t+1/2})\|_\infty \leq \lambda^{t+1} \|v_0 - T(v_0)\|_\infty$ **then**
        Set $v_{t+1} = v_{t+1/2}$
    else
        Set $v_{t+1} = T(v_t)$
    end if
end for

Algorithm 4 Value Iteration

input:
Initialize $V_0 = \mathbf{0}$ where $\mathbf{0}$
for $t \in \{0, 1, ..., T-1\}$ **do**
    $V_{t+1} = T(V_t)$
end for

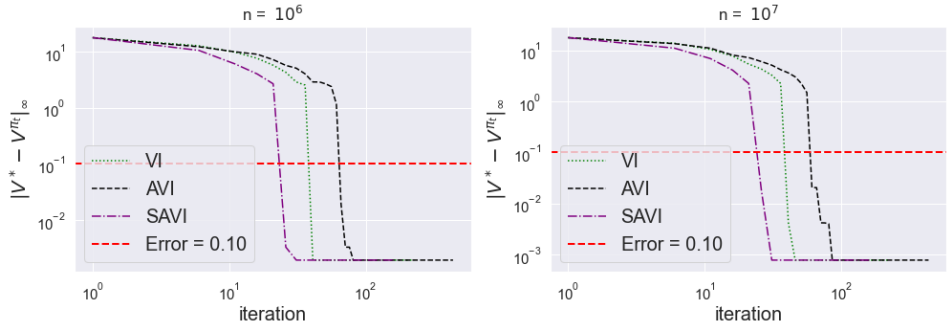


(a) $\gamma = 0.95$

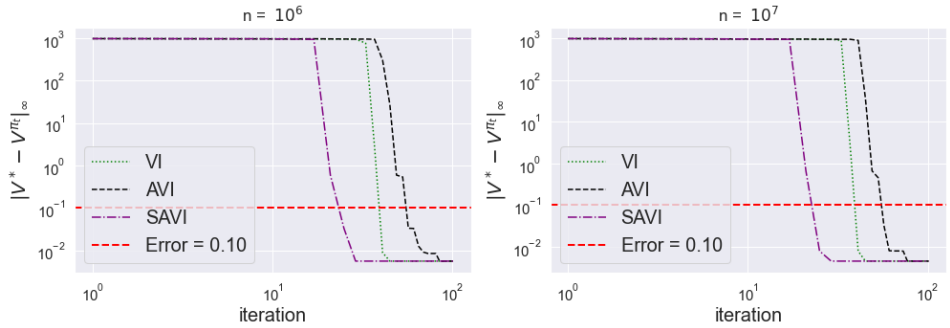


(b) $\gamma = 0.999$

Figure 4: Experiment for some more values of number of samples



(a) $\gamma = 0.95$



(b) $\gamma = 0.999$

Figure 5: Experiment for some more values of number of samples