Proceedings of Machine Learning Research vol 336:1–31, 2026 39th Annual Conference on Learning Theory

Strongly Polynomial Time Complexity of Policy Iteration for L_∞ Robust MDPs

Ali Asadi ALI.ASADI@IST.AC.AT
Krishnendu Chatterjee KRISHNENDU.CHATTERJEE@IST.AC.AT
Ehsan Goharshady EHSAN.GOHARSHADY@IST.AC.AT
Mehrdad Karrabi MEHRDAD.KARRABI@IST.AC.AT
Alipasha Montaseri ALIPASHA.MONTASERI@IST.AC.AT
Carlo Pagano CARLO.PAGANO@CONCORDIA.CA

Editors: Steve Hanneke and Tor Lattimore

Abstract

Markov decision processes (MDPs) are a fundamental model in sequential decision making. Robust MDPs (RMDPs) extend this framework by allowing uncertainty in transition probabilities and optimizing against the worst-case realization of that uncertainty. In particular, (s, a)-rectangular RMDPs with L_∞ uncertainty sets form a fundamental and expressive model: they subsume classical MDPs and turn-based stochastic games. We consider this model with discounted payoffs. The existence of polynomial and strongly-polynomial time algorithms is a fundamental problem for these optimization models. For MDPs, linear programming yields polynomial-time algorithms for any arbitrary discount factor, and the seminal work of Ye established strongly-polynomial time for a fixed discount factor. The generalization of such results to RMDPs has remained an important open problem. In this work, we show that a robust policy iteration algorithm runs in strongly-polynomial time for (s, a)-rectangular L_∞ RMDPs with a constant (fixed) discount factor, resolving an important algorithmic question.

Keywords: Robust Markov decision processes (RMDPs), Robust policy iteration, Strongly polynomial algorithms, Discounted-sum objectives

1. Introduction

Robust Markov Decision Processes. Markov decision processes (MDPs) are a fundamental model in sequential decision making, where an agent interacts with a finite-state stochastic environment (Puterman, 1994). Solving MDPs focuses on minimizing the expected payoff with respect to some given cost function. In the context of MDPs, a common assumption is that the transition function is known. Although this assumption is plausible from a theoretical perspective, it is not always justified in practice since MDPs are constructed from data, and the transition functions are estimated with a level of uncertainty. This issue has motivated the study of *Robust Markov decision processes* (RMDPs) (Nilim and El Ghaoui, 2005; Iyengar, 2005). RMDPs weaken the assumption by only assuming knowledge of some uncertainty set containing the true transition function. The goal of solving RMDPs is to minimize the worst-case expected payoff with respect to all possible choices of transition functions belonging to the uncertainty set.

Discounted-sum payoff. The agent aims to minimize a *payoff function* which formally captures the desired behavior of the model. Discounted-sum payoffs are the most fundamental payoffs, i.e., given a cost function and discount factor, the payoff of a run (infinite trajectory of states and actions) is the sum of discounted costs of the run.

Policies and values. Policies are recipes that define the choice of actions of the agent and the environment. They are functions that, given a finite sequence of states and actions (also known as history), return a distribution over actions. Given an RMDP and a discount factor, the payoff minimization is achieved over all possible choices of agent policies. The value of the agent at a state is the minimal payoff that the agent can guarantee against all policies of the environment.

$\mathbf{L}_\infty$ uncertainty sets. A central modeling choice in RMDPs is how we describe the uncertainty of transition probabilities. The literature considers several classes of uncertainty, including (s, a)-rectangular, s-rectangular, and more general non-rectangular models. In (s, a)-rectangular RMDPs, the environment selects transition probabilities separately for each state-action pair. In this paper, we adopt this rectangularity assumption for two reasons: (a) it matches a common data-driven setting where each state-action pair (s, a) has transition probabilities estimated from local samples, so uncertainty decomposes across state-action pairs; and (b) it preserves a dynamic-programming structure, yielding robust Bellman operators (Nilim and El Ghaoui, 2005; Iyengar, 2005). We further assume $\mathbf{L}_\infty$ uncertainty sets, i.e., for every state-action pair (s, a), the true transition vector lies in an $\mathbf{L}_\infty$ ball of radius $\delta_{s,a}$ around a nominal estimate $\hat{P}_{s,a}$.

Motivation. The study of (s, a)-rectangular RMDPs with $\mathbf{L}_\infty$ uncertainty sets is motivated by three reasons: (a) they generalize several classical formalisms, including MDPs, and turn-based zero-sum stochastic games, where two adversarial players interact over a finite-state graph (see reduction in Appendix A); (b) they provide a simple and interpretable bound on the *maximum coordinate-wise* estimation error for the transition probabilities, which arises naturally from sample-based estimation (Nilim and El Ghaoui, 2005; Givan et al., 2000; Delgado et al., 2016); and (c) they keep the robust Bellman update computationally efficient (Behzadian et al., 2021). Hence, (s, a)-rectangular RMDPs with $\mathbf{L}_\infty$ uncertainty sets are a fundamental model.

Exact vs. approximate value. While approximation algorithms for RMDPs are well-studied, computing the exact value is important for the following reasons. First, the exact value is a fundamental theoretical question: in MDPs with discounted-sum payoffs, strongly-polynomial additive approximation is easy via value iteration, and the seminal work of Ye (2011) establishes a strongly-polynomial algorithm for the exact value. Second, for MDPs and RMDPs with discounted-sum objectives, additive approximation guarantees are easy to obtain via value iteration, however, multiplicative-factor approximation (guaranteeing a high fraction of the optimal value) is not easy, and solving for the exact value gives a stronger guarantee.

Previous computational results. In the study of MDPs and related optimization problems, the study of efficient algorithms is a central problem. For example, MDPs with discounted-sum objectives can be solved in polynomial-time via linear-programming (d'Epenoux, 1963; Derman, 1972). It has been a long-standing problem to obtain a strongly polynomial-time algorithm where the number of arithmetic operations is polynomial independent of the bit length. For the special case of fixed discount factor, the seminal work of Ye (2011) obtains the first strongly-polynomial time algorithm, which has also been extended to stochastic games by Hansen et al. (2013). Efficient (i.e., polynomial and strongly-polynomial time) algorithms for RMDPs with $\mathbf{L}_\infty$ uncertainty sets is a fundamental algorithmic problem. The only claimed result in the literature is a polynomial-time bound from Behzadian et al. (2021) (which claims without proof that this follows from Hansen et al. (2013)). However, this claim was an oversight as the mentioned technique does not yield the desired result for two reasons. First, Hansen et al. (2013) studies turn-based stochastic games, and the only known reduction, presented in Chatterjee et al. (2024), is from (s, a)-*rectangular* RMDPs (not s-rectangular RMDPs) to turn-based stochastic games. Second, even for (s, a)-rectangular RMDPs, the reduction is not of polynomial size in general. For each uncertainty set, it introduces an action for each corner of the corresponding polytope, and for $\mathbf{L}_\infty$ uncertainty sets, the number of actions can be exponential.

Main open problem. Hence, the existence of polynomial and more importantly strongly-polynomial time algorithms for RMDPs with $\mathbf{L}_\infty$ uncertainty sets and a fixed discount factor is an important open question in the study of sequential decision making under uncertainty.

Our contributions. In this work, we affirmatively answer the above open problems. In particular, we show that a robust policy iteration algorithm terminates in strongly polynomial time, which is a generalization of the result of Hansen et al. (2013).

Technical contributions.

- We first consider the robust policy iteration algorithm on robust Markov chains (`RMC-PI`, Algorithm 1). These are RMDPs where the agent has a single action in every state.
  - Firstly, we introduce a novel potential function that tracks the effects of changing the optimal policy within the uncertainty set with respect to probability mass transfers allowed by the homotopy algorithm (Algorithm 2).
  - We then show bounds to relate the policy values and the defined potential function. Moreover, we prove Lemma 8, which is a novel combinatorial result over the number of most significant bits in the binary representation of unitary signed subset sums of a finite set of real numbers. We derive Lemma 8 as a consequence of the more general Theorem 20 (See Appendix H). We present here also the effective strengthening given by Theorem 21, which was subsequently and independently obtained, shortly after we established our Theorem 20, by a custom mathematics research agent, named *Aletheia*[1], and built upon Gemini Deep Think at Google DeepMind under the lead of Tony Feng. Although our own Theorem 20 is sufficient for Lemma 8, we decided to report the stronger version as in Theorem 21, both for its own intrinsic mathematical interest as well as to keep track of a contribution of AI to mathematics.
  - Finally, by the end of Section 4, we use the combinatorial result together with proved bounds to show that `RMC-PI` runs in strongly polynomial time given an RMC with n states and a constant discount factor.
- Secondly, we focus on the robust policy iteration algorithm on RMDPs (`RMDP-PI`) and deploy a similar pipeline as above to show that `RMDP-PI` runs in strongly polynomial time given an RMDP with n states, m actions, and a constant discount factor.

Related works. In this work we focus on the exact value computation for RMDPs with $\mathbf{L}_\infty$ uncertainty sets. There are many related works that consider a more general model, but for the value approximation problem. (s, a)-rectangular RMDPs with discounted-sum payoffs have been extensively studied in the literature (see Suilen et al. (2024)), starting with the seminal works of Nilim and El Ghaoui (2005); Iyengar (2005). In the general setting of (s, a)-rectangular uncertainty, both Nilim and El Ghaoui (2005); Iyengar (2005) presented *robust value iteration* methods to compute the approximate value. Moreover, Iyengar (2005) introduced a *robust policy iteration* algorithm to approximate the value of the same (s, a)-rectangular model. Later, Kaufman and Schaefer (2013) studied another variant of robust policy iteration for discounted-sum (s, a)-rectangular RMDPs and

1. *Aletheia*Feng et al. (2026) is a custom mathematics research agent built upon Gemini Deep Think, details at https://github.com/google-deepmind/superhuman/tree/main/aletheia.

showed that robust policy iteration converges to an approximate value. However, these works do not present polynomial or strongly-polynomial bounds for the exact value computation, which is the focus of this work.

2. Preliminaries and Model Description

Notation. For arbitrary set S, we use $\Delta(S)$ to show the set of all probability distributions over S, and use 2^S to show the set of all subsets of S. Moreover, we use $size(n) = \lceil \log_2 n \rceil$ and $size(\frac{p}{q}) = size(p) + size(q)$ as the bit-size of integer n and rational number $\frac{p}{q}$.

Markov Chains. A Markov chain (MC) is a tuple $M = (\mathcal{S}, c, P)$ where $\mathcal{S} = \{1, \dots, |\mathcal{S}|\}$ is the set of states, $c\colon \mathcal{S} \to \mathbb{R}$ specifies the cost of visiting each state, and $P\colon \mathcal{S} \to \Delta(\mathcal{S})$ denotes the transition probabilities. The transition probability function P can be seen as an $n \times n$ matrix where $P_{s,t} = P(s)[t]$. Similarly, the cost function c can be presented as a vector in $\mathbb{R}^n$. The semantics of MCs are defined over sequences $\pi = \langle \pi_0, \pi_1, \dots \rangle$ of states. Using $\Pr[\pi_{i+1}|\pi_i] = P_{\pi_i,\pi_{i+1}}$, the cylinder construction of Baier and Katoen (2008) yields a probability distribution over the set of all infinite sequences of states. Given a discount factor γ, the value function $\boldsymbol{v}_M(s)$ for each state $s \in \mathcal{S}$ is defined as the expected discounted total cost of sequences that start at s:

$$\boldsymbol{v}_M(s) = \mathbb{E}_\pi \big[\sum_i \gamma^i c(\pi_i) | \pi_0 = s \big] = \big[(\mathbb{I} - \gamma P)^{-1} c \big](s) \tag{1}$$

Robust MCs. Robust MCs (RMCs) are an extension of MCs by considering a particular uncertainty in the transition probabilities. Formally, an RMC is a tuple $\mathcal{M} = (\mathcal{S}, c, \mathit{Succ}, \mathcal{P})$ where $\mathcal{S}$ and c are the same as in MCs, $\mathit{Succ}\colon \mathcal{S} \to 2^{\mathcal{S}}$ indicates the set of possible successors of each state, and $\mathcal{P}\colon \mathcal{S} \to 2^{\Delta(\mathcal{S})}$ is the uncertain transition function satisfying $\mathcal{P}(s) \subseteq \Delta(\mathit{Succ}(s))$. The semantics of RMCs are defined with respect to environment policies. A stationary environment policy is a function $\tau\colon \mathcal{S} \to \Delta(\mathcal{S})$ mapping each state to an element of its uncertainty set. Formally, $\tau(s) \in \mathcal{P}(s)$ for each $s \in \mathcal{S}$. By fixing an environment policy τ, the uncertainty of the RMC is resolved, yielding an MC $\mathcal{M}^\tau = (\mathcal{S}, c, \mathcal{P}^\tau)$. The value of a policy τ is then defined as the value of its induced MC, i.e. $\boldsymbol{v}^\tau_{\mathcal{M}}(s) = \boldsymbol{v}_{\mathcal{M}^\tau}(s)$. The goal of the environment is to choose a policy that maximizes the expected discounted cost. Formally, an optimal policy τ^* is a policy that satisfies $\boldsymbol{v}^{\tau^*}_{\mathcal{M}}(s) = \max_\tau \boldsymbol{v}^\tau_{\mathcal{M}}(s)$. We drop the subscript $\mathcal{M}$ whenever it is clear from the context and use $\boldsymbol{v}^*$ to denote $\boldsymbol{v}^{\tau^*}$.

Robust Markov Decision Processes. It is possible to further extend RMCs into robust Markov Decision Processes (RMDPs) by considering a finite action space for the agent. Formally, an RMDP is a tuple $\mathcal{R} = (\mathcal{S}, c, \mathcal{A}, \mathit{Succ}, \mathcal{P})$ where $\mathcal{S}$ and c are defined similarly as before, $\mathcal{A}$ is a finite set of actions, $\mathit{Succ}\colon \mathcal{S} \times \mathcal{A} \to 2^{\mathcal{S}}$, and $\mathcal{P}\colon \mathcal{S} \times \mathcal{A} \to 2^{\Delta(\mathcal{S})}$ is the uncertain transition function where $\mathcal{P}(s,a) \subseteq \Delta(\mathit{Succ}(s,a))$ for each state-action pair (s,a). The semantics of RMDPs are defined with respect to an agent policy $\sigma\colon \mathcal{S} \to \Delta(\mathcal{A})$ that maps each state to a distribution over the action-set $\mathcal{A}$. By fixing such an agent policy, the agent's choices are resolved, leaving an RMC $\mathcal{R}^\sigma = (\mathcal{S}, c, \mathit{Succ}^\sigma, \mathcal{P}^\sigma)$. The value of an agent policy is defined as the value of its induced RMC, i.e. $\boldsymbol{v}^\sigma_{\mathcal{R}}(s) = \boldsymbol{v}_{\mathcal{R}^\sigma}(s)$. The agent's objective is to find a policy that minimizes the worst-case expected cost. The optimal agent policy is then denoted by σ^* and satisfies:

$$\boldsymbol{v}^*_{\mathcal{R}}(s) = \boldsymbol{v}^{\sigma^*}_{\mathcal{R}}(s) = \min_\sigma \boldsymbol{v}^*_{\mathcal{R}^\sigma}(s) = \min_\sigma \max_\tau \boldsymbol{v}_{\mathcal{R}^{\sigma,\tau}}(s) = \min_\sigma \max_\tau \big[(\mathbb{I} - \gamma \mathcal{P}^{\sigma,\tau})^{-1} c \big](s) \tag{2}$$

When $\mathcal{R}$ is clear from the context, we use $\boldsymbol{v}^{\sigma}(s)$ and $\boldsymbol{v}^{\sigma,\tau}(s)$ to denote $\boldsymbol{v}_{\mathcal{R}^{\sigma}}(s)$ and $\boldsymbol{v}_{\mathcal{R}^{\sigma,\tau}}(s)$, respectively, for agent policy σ and environment policy τ. We use σ^* and τ^* for the optimal agent policy and the best environment response, respectively, satisfying $\boldsymbol{v}^*_{\mathcal{R}}(s) = \boldsymbol{v}^{\sigma^*}(s) = \boldsymbol{v}^{\sigma^*,\tau^*}(s)$.

Assumption 1. (Rectangularity) The uncertainty considered in this paper is also known as (s,a)-rectangular uncertainty since the environment can observe the agent's action before choosing a transition function from the uncertainty set. Other classes of uncertainty, such as s-rectangularity, have been studied in the literature (Suilen et al. (2024)) but are not the focus of our work.

Assumption 2. ($\mathbf{L}_\infty$ Uncertainty Sets) In this work, we focus on RMCs and RMDPs whose uncertainty sets are defined as $\mathbf{L}_\infty$ balls around a nominal transition function. Formally, in an L_∞ RMC $\mathcal{M}$, for each $s \in \mathcal{S}$, there exists non-negative real number δ_s and a nominal probabilistic transition function $\hat{P}_s \in \Delta(\mathcal{S})$ such that $\mathcal{P}(s) = \{P \in \Delta(\mathcal{S}) | \|P - \hat{P}_s\|_\infty \leq \delta_s\}$. Similarly, in an L_∞ RMDP, for each $s \in \mathcal{S}$ and $a \in \mathcal{A}$, there exists a non-negative real number $\delta_{s,a}$ and nominal transition function $\hat{P}_{s,a}$ such that $\mathcal{P}(s,a) = \{P \in \Delta(\mathcal{S}) | \|P - \hat{P}_{s,a}\|_\infty \leq \delta_{s,a}\}$. For the rest of the paper, we assume all the uncertainty sets are of this type.

Policy Types. The policies considered in the above definitions are positional (their value depends only on the current state of the model). In general, policies can also depend on the history of visited states and chosen actions. However, it is a classical result that in discounted-sum RMDPs, positional (memoryless) deterministic optimal policies exist (Iyengar (2005); Nilim and El Ghaoui (2005)). Hence, in the rest of the paper, all the policies are assumed to be positional and deterministic, i.e. an agent policy in an RMDP is a map $\sigma \colon \mathcal{S} \to \mathcal{A}$ and an environment policy is a map $\tau \colon \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ where $\tau(s,a) \in \mathcal{P}(s,a)$.

Strongly Polynomial Algorithms. An algorithm is strongly polynomial if the following two conditions hold: (i) The number of arithmetic operations is bounded by a polynomial in the number of input variables and constraints, and (ii) the size of all numbers during the computation is bounded by a polynomial in the input size.

Recall that an algorithm runs in polynomial time if its runtime is polynomial in the input (bit-length) size. Hence, in polynomial time algorithms, the runtime of the algorithm might depend on the bit-size of the input coefficients. In contrast, in strongly polynomial algorithms, the number of arithmetic operations is independent of the numerical values of the coefficients, and any dependence on the input coefficients is limited to their unavoidable contribution to the input encoding length.

Problem Statement. Given an RMC $\mathcal{M}$ or an RMDP $\mathcal{R} = (\mathcal{S}, c, \mathcal{A}, \mathit{Succ}, \mathcal{P})$ that satisfies Assumptions 1 and 2, algorithms such as *value iteration* and *policy iteration* are designed for computing the value function $\boldsymbol{v}$ together with the optimal agent policy σ^* and environment policy τ^* that correspond to $\boldsymbol{v}$. Our goal in this work is to study the policy iteration method (explained in Section 3) and to show that it is strongly polynomial.

3. Overview: Robust Policy Iteration

We state our two main results, and prove them across Sections 4 and 5. Firstly, we show that policy iteration over RMCs (`RMC-PI`, Algorithm 1) runs in strongly polynomial time:

Theorem (Strongly Polynomial `RMC-PI`) *`RMC-PI` is strongly polynomial given an RMC $\mathcal{M}$ with n states and a constant discount factor γ.*

Algorithm 1 Policy Iteration For Robust Markov Chains (`RMC-PI`)

Input : RMC $\mathcal{M} = (\mathcal{S}, c, \mathit{Succ}, \mathcal{P})$, discount factor γ
Output: The optimal environment policy τ^*
$\tau^0 \leftarrow$ `arbitrary environment policy`
$t \leftarrow 0$
Repeat
  `// Policy Evaluation:`
  $\boldsymbol{v}^t \leftarrow (\mathbb{I} - \gamma \mathcal{P}^{\tau^t})^{-1} c$
  $t \leftarrow t+1$
  `// Policy Improvement via Homotopy Algorithm 2:`
  forall $s \in \mathcal{S}$ **do**
    $\tau^t(s) \leftarrow \underset{\mathbf{p} \in \mathcal{P}(s)}{\operatorname{argmax}}\, c(s) + \gamma \mathbf{p}^\top \boldsymbol{v}^{t-1}$
  end
Until $\tau^t = \tau^{t-1}$;
return τ^t

The above theorem is then utilized in Section 5 to show that policy iteration over RMDPs (`RMDP-PI`, Algorithm 3) also runs in strongly polynomial time:

Theorem (Strongly Polynomial `RMDP-PI`) *`RMDP-PI` is strongly polynomial given an RMDP $\mathcal{R}$ with n states, m actions and a constant discount factor γ.*

Policy Iteration. Policy iteration is a classical method for computing the value function in Markovian models (Puterman (1994)). Algorithm 1 shows the policy iteration algorithm for RMCs. Intuitively, given an RMC $\mathcal{M}$ and discount factor γ, the algorithm starts from an arbitrary initial environment policy τ^0. Then, in each iteration, it first evaluates the current policy by computing $\boldsymbol{v}^{\tau^t}$ and then greedily improves τ^t into τ^{t+1} in order to maximize the expected cost based on the computed value function.

The main difference between `RMC-PI` in Algorithm 1 and the classical policy iteration on Markov Decision Processes (Puterman (1994)) lies in the policy improvement step (also known as computing the Bellman operator), where in MDPs, the argmax is taken over the finite set of actions, but in RMCs it is taken over the possibly *infinite* uncertainty set $\mathcal{P}$.

Homotopy Algorithm. Given that $\mathcal{P}$ is an $\mathbf{L}_\infty$ uncertainty set, the policy improvement can be computed using linear programming. However, solving an LP instance is inefficient, its runtime being $O(n^{3.5}L)$, which depends on the bit precision L of the input system. In Behzadian et al. (2021), authors present a *homotopy* algorithm for computing the policy improvement, which runs in $O(n \log n)$. Algorithm 2 illustrates their approach adapted to our notations.

Intuitively, given the RMC $\mathcal{M}$, state $s \in \mathcal{S}$, and the current policy's value vector $\boldsymbol{v}$, the algorithm first sorts the successor states with respect to their $\boldsymbol{v}$ value so that $\boldsymbol{v}(s_1) \geq \boldsymbol{v}(s_2) \geq \cdots \geq \boldsymbol{v}(s_{|\mathit{Succ}(s)|})$. It then proceeds by adapting a two-pointer technique for increasing the probability of transitioning to states with higher values while decreasing the probability of transitioning to states with lower values. This is done by keeping two indices hi and lo. While $hi < lo$, the algorithm increases $\mathbf{p}_{s_{hi}}$ and decreases $\mathbf{p}_{s_{lo}}$ as much as possible. If $\mathbf{p}_{s_{hi}}$ could not be increased further, i.e. by reaching 1 or $\hat{P}_{s,s_{hi}} + \delta(s)$, then hi is increased by one. Similarly, if $\mathbf{p}_{s_{lo}}$ could not be decreased

Algorithm 2 Homotopy Algorithm for Policy Improvement

Input : RMC $\mathcal{M} = (\mathcal{S}, c, \mathit{Succ}, \mathcal{P})$, successor set $\mathit{Succ}(s)$, nominal transition matrix $\hat{P}_s$, state $s \in \mathcal{S}$, radius $\delta(s)$ for the uncertainty set, value vector $\boldsymbol{v}$

Output: Distribution $\mathbf{p} \in \mathcal{P}(s)$ maximizing $\mathbf{p}^\top \boldsymbol{v}$

$\mathcal{S}' \leftarrow \mathit{Succ}(s)$

Sort states $s_1, \dots, s_{|\mathcal{S}'|}$ of $\mathcal{S}'$ so that $\boldsymbol{v}(s_1) \geq \boldsymbol{v}(s_2) \geq \cdots \geq \boldsymbol{v}(s_{|\mathcal{S}'|})$

$\mathbf{p} \leftarrow \hat{P}_s$

$hi \leftarrow 1, lo \leftarrow |\mathcal{S}'|$

$b_{hi} \leftarrow \delta(s), \quad b_{lo} \leftarrow \delta(s)$

while $hi < lo$ **do**

    $d_{hi} \leftarrow \min(b_{hi},\, 1 - \mathbf{p}_{s_{hi}})$

    $d_{lo} \leftarrow \min(b_{lo},\, \mathbf{p}_{s_{lo}})$

    $t \leftarrow \min(d_{hi},\, d_{lo})$

    $\mathbf{p}_{s_{hi}} \leftarrow \mathbf{p}_{s_{hi}} + t$

    $\mathbf{p}_{s_{lo}} \leftarrow \mathbf{p}_{s_{lo}} - t$

    $b_{hi} \leftarrow b_{hi} - t$

    $b_{lo} \leftarrow b_{lo} - t$

    if $b_{hi} = 0$ ***or*** $\mathbf{p}_{s_{hi}} = 1$ **then**

        $hi \leftarrow hi + 1 \;\; b_{hi} \leftarrow \delta(s)$

    else

        $lo \leftarrow lo - 1 \;\; b_{lo} \leftarrow \delta(s)$

    end

end

return $\mathbf{p}$

further, i.e. by reaching 0 or $\hat{P}_{s,s_{lo}} - \delta(s)$, then lo is decreased by one. See Behzadian et al. (2021) for correctness arguments.

Properties of Homotopic Policies. The distribution $\mathbf{p}$ generated by Algorithm 2 possesses a specific structure that becomes useful later in our analysis. Essentially, there are four disjoint sets of states $R_{\mathbf{p}}$, $D_{\mathbf{p}}$, $Z_{\mathbf{p}}$ and $I_{\mathbf{p}}$, where:

- for $s' \in R_{\mathbf{p}}$, $\mathbf{p}_{s'} = \min(\hat{P}_{s,s'} + \delta(s), 1)$,
- for $s' \in D_{\mathbf{p}}$, $\mathbf{p}_{s'} = \hat{P}_{s,s'} - \delta(s)$,
- for $s' \in Z_{\mathbf{p}}$, $\hat{P}_{s,s'} \leq \delta(s) \;\wedge\; \mathbf{p}_{s'} = 0$, and
- $|I_{\mathbf{p}}| \leq 1$ and for $s' \in I_{\mathbf{p}}$, $\hat{P}_{s,s'} - \delta(s) < \mathbf{p}_{s'} = 1 - \sum_{s'' \neq s'} \mathbf{p}_{s''} < \hat{P}_{s,s'} + \delta(s)$.

Intuitively, $R_{\mathbf{p}}$ states are *receiving* $\delta(s)$ probability mass. This mass is *donated* by donor states in $D_{\mathbf{p}}$ and $Z_{\mathbf{p}}$. Moreover, there is at most one state in $I_{\mathbf{p}}$ that receives/donates *incompletely*. These properties immediately show that there are at most exponentially $\mathcal{O}(n \cdot 3^n)$ many different homotopic policies that Algorithm 2 can choose from.

Strongly Polynomial Runtime of `RMC-PI`. Our primary objective is to demonstrate that the policy iteration method in Algorithm 1 runs in strongly polynomial time. Specifically, this means the number of policy improvement iterations is bounded by a polynomial independent of the cost function, nominal probabilities, or uncertainty set radii. We formally prove this result in Section 4.

Algorithm 3 Policy Iteration For Robust Markov Decision Process `RMDP-PI`

Input : RMDP $\mathcal{R} = (\mathcal{S}, c, \mathcal{A}, Succ, \mathcal{P})$, discount factor γ
Output: The optimal policy σ^*
$\sigma^0 \leftarrow$ `arbitrary agent policy`
$t \leftarrow 0$
Repeat
  `// Policy Evaluation, Invoke Algorithm 1:`
  $\tau^t \leftarrow \texttt{RMC-PI}(\mathcal{R}^{\sigma^t}, \gamma)$
  $\boldsymbol{v}^t \leftarrow (I - \gamma \mathcal{P}^{\sigma^t, \tau^t})^{-1} c$
  $t \leftarrow t + 1$
  `// Policy Improvement:`
  forall $s \in \mathcal{S}$ **do**
    $\sigma^t(s) \leftarrow \underset{a \in \mathcal{A}}{\operatorname{argmin}} \max_{\mathbf{p} \in \mathcal{P}(s,a)} c(s) + \gamma \mathbf{p}^\top \boldsymbol{v}^{t-1}$
  end
Until $\sigma^t = \sigma^{t-1}$;
return σ^t

Policy Iteration for RMDPs. We can extend policy iteration to compute optimal policies for RMDPs, utilizing Algorithm 1 as a key component. Algorithm 3 outlines this process. The method begins with an arbitrary agent policy, σ^0. In the t-th iteration, it first computes the optimal environment counter-policy, τ^t, against σ^t (using Algorithm 1) and evaluates the value function $\boldsymbol{v}^t$ for $\mathcal{R}^{\sigma^t, \tau^t}$. Finally, it uses $\boldsymbol{v}^t$ to improve the agent policy σ^t greedily.

Strongly Polynomial Runtime of `RMDP-PI`. Our final objective is to establish a strongly polynomial upper bound on the runtime of the `RMDP-PI` algorithm. In this context, the bound depends on the number of states, the available actions, and the discount factor. As with the RMC case, we formally prove this result in Section 5.

4. Robust Markov Chains

In this section, we focus our attention on RMCs and their corresponding policy iteration algorithm `RMC-PI` presented in Algorithm 1. Specifically, we assume `RMC-PI` is executed on an RMC $\mathcal{M} = (\mathcal{S}, c, Succ, \mathcal{P})$ with an $\mathbf{L}_\infty$ uncertainty set, $|\mathcal{S}| = n$ and a constant discount factor γ.

Let $\mathcal{M}$ be the input RMC in an execution of `RMC-PI`. The Bellman operator for computing the optimal values in $\mathcal{M}$ is defined as:

$$(\mathcal{T}\boldsymbol{v})_s = \mathbf{c}_s + \gamma \max_{\mathbf{p} \in \mathcal{P}(s)} \mathbf{p}^\top \boldsymbol{v}. \tag{3}$$

We first note that `RMC-PI` converges monotonically to the unique fixed-point of $\mathcal{T}$ at an exponential rate. This is a classical result in the analysis of Markov models (Puterman (1994); Iyengar (2005)), presented in Lemma 1 (proved in Appendix B) and Lemma 2 (proved in Appendix C).

Lemma 1 *The following statements hold:*

- *For every $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^n$, $\|\mathcal{T}\boldsymbol{u} - \mathcal{T}\boldsymbol{v}\|_\infty \leq \gamma \|\boldsymbol{u} - \boldsymbol{v}\|_\infty$.*
- *For every environment policy τ with value vector $\boldsymbol{v}^\tau$, $\mathcal{T}\boldsymbol{v}^\tau \succcurlyeq \boldsymbol{v}^\tau$.*

- *There exists a unique vector* $\boldsymbol{v}^* \in \mathbb{R}^n$ *such that* $\mathcal{T}\boldsymbol{v}^* = \boldsymbol{v}^*$.

Using the improvement properties established in Lemma 1, we show that the sequence of values generated by `RMC-PI` is non-decreasing and converging at an exponential rate:

Lemma 2 *Let* $\tau^0, \tau^1, \ldots$ *be the sequence of policies generated by* `RMC-PI` *(Algorithm 1), and let* $\boldsymbol{v}^t$ *be the value of* τ^t*. Then the following statements hold for each* t*:*

- $\boldsymbol{v}^{t+1} \succcurlyeq \boldsymbol{v}^t$
- $\left\|\boldsymbol{v}^t - \boldsymbol{v}^*\right\|_\infty \leq \gamma^t \left\|\boldsymbol{v}^0 - \boldsymbol{v}^*\right\|_\infty$

Lemma 2 establishes that policy iteration converges exponentially fast to the optimal policy. Given the exponential upper-bound on the number of homotopic policies that Algorithm 2 can return, Lemma 2 immediately implies that `RMC-PI` terminates in finite time and computes the optimal environment policy τ^* together with its value function $\boldsymbol{v}^*$.

To show that `RMC-PI` is strongly polynomial, we must show two statements: (i) it terminates within a polynomial number of arithmetic operations, and (ii) the bit-size of all intermediate values remains polynomial in the input size. Both conditions are satisfied if the total number of iterations of `RMC-PI` is bounded by a polynomial in $|\mathcal{S}|$. Specifically, the former is implied by the fact that both policy evaluation and policy improvement are themselves polynomial. The latter holds because each iteration recomputes $\boldsymbol{v}^t = (\mathbb{I} - \gamma \mathcal{P}^{\tau^t})^{-1}\mathbf{c}$ from scratch: the entries of $\mathcal{P}^{\tau^t}$ belong to a fixed finite set determined by $\hat{P}$ and δ (by the structure of τ^t), so their bit-size is polynomial in the input bit-size, and matrix inversion therefore yields a $\boldsymbol{v}^t$ whose bit-size is polynomial in the input bit-size and $size(\gamma)$, uniformly in t. So, we focus on showing that the number of iterations of `RMC-PI` is polynomial in $|\mathcal{S}|$.

To this end, our goal is to bound the change in $\left\|\boldsymbol{v}^{t+1} - \boldsymbol{v}^*\right\|_\infty$ compared to $\left\|\boldsymbol{v}^t - \boldsymbol{v}^*\right\|_\infty$ and then use the bound to obtain bounds on the maximum number of iterations before termination. Given the structure of the policies returned by the homotopy algorithm (Algorithm 2), we introduce a specific *potential function* $f_\tau(s, s', s'')$ to track how much the value of τ can be improved by donating some probability mass from $\mathcal{P}^\tau_{s,s''}$ to $\mathcal{P}^\tau_{s,s'}$:

Definition 3 (Potential Function) *Given a policy* τ*, for any* $s, s', s'' \in \mathcal{S}$ *define:*

$$f_\tau(s, s', s'') := \min\left(\mathcal{P}^*_{s,s'} - \mathcal{P}^\tau_{s,s'},\ \mathcal{P}^\tau_{s,s''} - \mathcal{P}^*_{s,s''}\right)\left(\boldsymbol{v}^*_{s'} - \boldsymbol{v}^*_{s''}\right),$$

Intuitively, the $\min$ fragment in the definition of the potential function specifies the probability mass being donated to $\mathcal{P}^\tau_{s,s'}$ from $\mathcal{P}^\tau_{s,s''}$, and the $\boldsymbol{v}^*$ fragment quantifies the change in values. Note that $f_\tau(s, s', s'')$ can take negative values; the maximizing triple used in Lemmas 5 and 6 is, however, always non-negative (e.g., taking $s' = s''$ yields $f_\tau = 0$).

We now establish a lower bound on the optimality gap in terms of the potential function. This result (proved in Appendix D) connects the potential function to the actual difference in values.

Lemma 4 *For every policy* τ*, and states* $s, s', s'' \in \mathcal{S}$ *it holds that* $\boldsymbol{v}^*_s - \boldsymbol{v}^\tau_s \geq \gamma\, f_\tau(s, s', s'')$.

Next, we derive an upper-bound on the distance between $\boldsymbol{v}^\tau$ and $\boldsymbol{v}^*$ based on the potential of τ. We view the difference between the optimal transition matrix and the current policy's matrix as a set of probability mass transfers. By bounding the impact of these transfers using the potential function, we link the global error (distance to optimal $\boldsymbol{v}^*$) back to the local slack (potential function).

Lemma 5 *Let τ be a policy. Let $s, s', s'' \in \mathcal{S}$ be the states that maximize $f_\tau(s, s', s'')$. Then:*

$$\|\boldsymbol{v}^* - \boldsymbol{v}^\tau\|_\infty \leq \frac{n^2\gamma}{1-\gamma} f_\tau(s, s', s'').$$

Proof Sketch. We first show that $\|\boldsymbol{v}^* - \boldsymbol{v}^\tau\|_\infty \leq \gamma \left\|(\mathbb{I} - \gamma\mathcal{P}^\tau)^{-1}\right\|_\infty \|(\mathcal{P}^* - \mathcal{P}^\tau)\boldsymbol{v}^*\|_\infty$ and proceed by showing an upper-bound on each factor of the RHS. (i) We show: $\left\|(\mathbb{I} - \gamma\mathcal{P}^\tau)^{-1}\right\|_\infty \leq \frac{1}{1-\gamma}$ by using the geometric series sum. (ii) We then show that $\|(\mathcal{P}^* - \mathcal{P}^\tau)\boldsymbol{v}^*\|_\infty \leq n^2 f_\tau(s, s', s'')$ by decomposing $\mathcal{P}^* - \mathcal{P}^\tau$ into a series of at most n^2 *mass transfer* operations each contributing at most $f_\tau(s, s', s'')$ to the LHS. The full proof is presented in Appendix E. ■

As a byproduct of Lemma 4 and Lemma 5, we can now obtain a bound between $\|\boldsymbol{v}^* - \boldsymbol{v}^\tau\|_\infty$ and $\left\|\boldsymbol{v}^* - \boldsymbol{v}^{\tau'}\right\|_\infty$ based on the potential functions of τ and τ':

Lemma 6 *Assume τ and τ' are two policies. Let $s, s', s'' \in \mathcal{S}$ be the states that maximize $f_\tau(s, s', s'')$. Suppose further that τ' satisfies:*

$$\min\left(\mathcal{P}^*_{s,s'} - \mathcal{P}^{\tau'}_{s,s'},\ \mathcal{P}^{\tau'}_{s,s''} - \mathcal{P}^*_{s,s''}\right) \geq \frac{1}{2}\min\left(\mathcal{P}^*_{s,s'} - \mathcal{P}^{\tau}_{s,s'},\ \mathcal{P}^{\tau}_{s,s''} - \mathcal{P}^*_{s,s''}\right).$$

Then $\left\|\boldsymbol{v}^* - \boldsymbol{v}^{\tau'}\right\|_\infty \geq \frac{1-\gamma}{2n^2}\|\boldsymbol{v}^* - \boldsymbol{v}^\tau\|_\infty$.

Proof Sketch. We chain the upper bound of Lemma 5 (for τ) and the lower bound of Lemma 4 (for τ') to derive the result. Full proof is available in Appendix F. ■

We now show that the algorithm cannot get "stuck" making small updates to the same constraints. Specifically, we prove that within a fixed window of L steps, the difference between the probability at $\mathcal{P}^*$ and $\mathcal{P}^\tau$ on the critical constraint must be at least halved. The intuition relies on a contradiction: if this difference remained large, Lemma 6 would imply that the value error also remains large. However, we know from Lemma 2 that the value error contracts exponentially. Therefore, this difference must decrease significantly to match this contraction.

Lemma 7 *Let $\tau^0, \ldots, \tau^T$ be the sequence of policies generated by the policy iteration algorithm. Define the step threshold $L := \log_\gamma\left(\frac{1-\gamma}{2n^2}\right)$. For a specific policy τ^t, let $(s, s', s'') = \operatorname{argmax}_{s,s',s''\in\mathcal{S}} f_{\tau^t}(s, s', s'')$ be the triple with the maximum potential in τ^t. Then, in any subsequent policy τ^l with $l > t + L$, we have:*

$$\min\left(\mathcal{P}^*_{s,s'} - \mathcal{P}^{\tau^l}_{s,s'},\ \mathcal{P}^{\tau^l}_{s,s''} - \mathcal{P}^*_{s,s''}\right) \leq \frac{1}{2}\min\left(\mathcal{P}^*_{s,s'} - \mathcal{P}^{\tau^t}_{s,s'},\ \mathcal{P}^{\tau^t}_{s,s''} - \mathcal{P}^*_{s,s''}\right).$$

Lemma 7 (proved in Appendix G) shows that after every L iterations of `RMC-PI`, the amount of $\min(\mathcal{P}^*_{s,s'} - \mathcal{P}^{\tau^l}_{s,s'}, \mathcal{P}^{\tau^l}_{s,s''} - \mathcal{P}^*_{s,s''})$ reduces by at least half. This means that the *most significant bit (MSB)* in its binary presentation changes by at least one index. Our goal is to show that there are polynomially many possible MSBs for this value, hence it can only change a polynomial amount of times. To this end, we introduce the following combinatorial lemma (proved in Appendix H):

Lemma 8 *Let c be a constant positive integer. Let X be a finite set of nonnegative real numbers. Define the set of all signed subset sums as:*

$$A(X) = \left\{ \left| \sum_{x \in X} g(x) \cdot x \right| \;\middle|\; g : X \to \{-c, \cdots, 0, \cdots, c\} \right\}.$$

Define the degree *of a set of positive real numbers $Y \subseteq \mathbb{R}_{>0}$ as the number of distinct MSBs in Y:* $\mathrm{Deg}(Y) = |\{\lfloor \log_2 y \rfloor \mid y \in Y\}|$ *. Then for all X, it holds that*

$$\mathrm{Deg}(A(X) \setminus \{0\}) \in \mathcal{O}(|X| \log |X|).$$

Lemma 8 shows that for a set X of non-negative reals, the number of MSBs in the set $A(X)$ is a polynomial bounded by a polynomial in the size of X. We derive the final complexity bound by combining the geometric contraction of the discrepancy with the finite number of possible discrepancy values.

Theorem 9 *`RMC-PI` terminates with an optimal policy in $\mathcal{O}\left(n^4 \log n \cdot \frac{\log\left(\frac{1-\gamma}{n^2}\right)}{\log \gamma}\right)$ iterations where $n = |\mathcal{S}|$.*

Proof Sketch. The proof is done in three steps:

- **Geometric Contraction:** Lemma 7 shows that for the maximizing triple, the discrepancy between the current and optimal transition function decreases by a factor of at least two every L steps.
- **Combinatorial Finiteness:** Lemma 8 restricts the possible values of the transition probabilities. These values are generated by a finite set of parameters (the nominal probabilities and uncertainty radii). Specifically, by invoking Lemma 8 on a set with $O(n)$ elements, we show that there are $O(n \log n)$ possible MSBs for the discrepancy.
- There are at most n^3 possibilities for the triple with the highest discrepancy. Hence, after every $O(n^3 \cdot L)$ iterations, the MSB of the highest discrepancy decreases by at least one. Given that there are $O(n \log n)$ possible different MSBs, the algorithm must terminate within $O(n^4 \log n \cdot L)$ iterations. Dropping constants gives the bound in the theorem statement.

Full proof is available in Appendix I. ■

Since the number of iterations required by the `RMC-PI` algorithm (Algorithm 1) is polynomial, we can conclude that `RMC-PI` is a strongly polynomial algorithm.

Corollary 10 *`RMC-PI` is strongly polynomial given an RMC with n states and constant discount factor γ.*

5. Robust Markov Decision Processes

In this section, we analyze `RMDP-PI` (Algorithm 3) and its convergence to the optimal policy. Let $\mathcal{R} = (\mathcal{S}, c, \mathcal{A}, \mathit{Succ}, \mathcal{P})$ be the input RMDP where $|\mathcal{S}| = n$, $|\mathcal{A}| = m$ and $\mathcal{P}$ is an $\mathbf{L}_\infty$ uncertainty set. We follow a pipeline similar to our analysis of `RMC-PI` in Section 4: We first recall the Bellman operator and its properties (Lemma 11), which imply the exponential convergence rate of `RMDP-PI` (Lemma 12). We then define a potential function and show a lower and an upper bound on the value error of policies based on their potentials (Lemma 14). Finally, we use the obtained bounds to show that `RMDP-PI` terminates after a polynomial number of iterations (Theorem 17).

The Bellman optimality operator $\mathcal{T}$ for $\mathcal{R}$ is defined as:

$$(\mathcal{T}\boldsymbol{v})_s = \mathbf{c}_s + \gamma \min_{a \in \mathcal{A}} \max_{\mathbf{p} \in \mathcal{P}_{s,a}} \mathbf{p}^\top \boldsymbol{v}.$$

We first establish that $\mathcal{T}$ is a contraction mapping with a unique fixed point. This result is standard in the literature of robust dynamic programming (Iyengar (2005)). The full proof of Lemma 11 is available in Appendix J.

Lemma 11 *The following statements hold:*

- *For all $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^n$, $\|\mathcal{T}\boldsymbol{u} - \mathcal{T}\boldsymbol{v}\|_\infty \leq \gamma \|\boldsymbol{u} - \boldsymbol{v}\|_\infty$.*
- *For every agent policy σ with value vector $\boldsymbol{v}^\sigma$, $\mathcal{T}\boldsymbol{v}^\sigma \preccurlyeq \boldsymbol{v}^\sigma$.*
- *There exists a unique $\boldsymbol{v}^* \in \mathbb{R}^n$ such that $\mathcal{T}\boldsymbol{v}^* = \boldsymbol{v}^*$.*

Next, we prove that the sequence of value functions generated by Policy Iteration is non-increasing and exponentially converging to the optimal value. This monotonicity ensures that the algorithm consistently improves the policy. The full proof of Lemma 12 is available in Appendix K.

Lemma 12 *Let σ^t and σ^{t+1} be two consecutive policies in the sequence generated by `RMDP-PI` (Algorithm 3) and let $\mathbf{v}^t$ be the value of σ^t. Then the following statements hold for each t:*

- $\boldsymbol{v}^{t+1} \preccurlyeq \boldsymbol{v}^t$
- $\left\|\boldsymbol{v}^t - \boldsymbol{v}^*\right\|_\infty \leq \gamma^t \left\|\boldsymbol{v}^0 - \boldsymbol{v}^*\right\|_\infty$

We proceed by defining the potential function. Intuitively, the potential of an action a in state s tracks the additional cost incurred by deviating from the optimal policy σ^* and choosing a at s instead of $\sigma^*(s)$.

Definition 13 (Potential function) *Given a state $s \in \mathcal{S}$ and an action $a \in \mathcal{A}$ the potential function $f(s,a)$ is defined as follows:*

$$f(s,a) := \max_{\mathbf{p} \in \mathcal{P}_{s,a}} \mathbf{p}^\top \boldsymbol{v}^* - \max_{\mathbf{p} \in \mathcal{P}_{s,\sigma^*(s)}} \mathbf{p}^\top \boldsymbol{v}^*$$

The function $f(s,a)$ is the cost-minimization analogue of the standard advantage function in MDPs, taken with respect to the optimal policy σ^*.

Next, similar to Lemma 4 and Lemma 5, we establish a lower-bound and an upper-bound on the difference of a policy value vector and the optimal value in terms of the potential function. The full proof of Lemma 14 is available in Appendix L.

Lemma 14 *Let σ be an arbitrary policy. The following holds:*

- ***(Lower-bound)*** *If $\sigma(s) = a$ for some state $s \in \mathcal{S}$, then $\boldsymbol{v}^\sigma_s - \boldsymbol{v}^*_s \geq \gamma f(s,a)$.*
- ***(Upper-bound)*** *Suppose $\hat{s} := \mathrm{argmax}_{s \in \mathcal{S}} f(s, \sigma(s))$. Then, $\|\boldsymbol{v}^\sigma - \boldsymbol{v}^*\|_\infty \leq \frac{\gamma}{1-\gamma} f(\hat{s}, \sigma(\hat{s}))$.*

We now combine the lower and upper bounds to derive a relationship between the value errors of two different policies. Specifically, we show that if the action chosen at the state with maximum potential remains unchanged between two policies, the value error cannot decrease by more than a factor of $1-\gamma$.

Lemma 15 *Let σ and σ' be two arbitrary policies. Let $\hat{s} = \operatorname{argmax}_{s \in \mathcal{S}} f(s, \sigma(s))$. Also assume that $\sigma'(\hat{s}) = \sigma(\hat{s})$. Then we have that:*

$$\left\| \boldsymbol{v}^{\sigma'} - \boldsymbol{v}^* \right\|_\infty \geq (1-\gamma) \cdot \| \boldsymbol{v}^\sigma - \boldsymbol{v}^* \|_\infty$$

Proof $\left\| \boldsymbol{v}^{\sigma'} - \boldsymbol{v}^* \right\|_\infty \geq (\boldsymbol{v}_{\hat{s}}^{\sigma'} - \boldsymbol{v}_{\hat{s}}^*) \geq \gamma f(\hat{s}, \sigma'(\hat{s})) = \gamma f(\hat{s}, \sigma(\hat{s})) \geq (1-\gamma) \| \boldsymbol{v}^\sigma - \boldsymbol{v}^* \|_\infty$ where the second and the last inequalities are due to the bounds in Lemma 14. ■

We now utilize the exponential convergence rate established in Lemma 12 to derive that an action with maximum potential at some iteration cannot persist for more than L iterations. The full proof of Lemma 16 is available in Appendix M.

Lemma 16 *Let $\sigma^0, \ldots, \sigma^T$ be the sequence of agent policies generated by `RMDP-PI`. Let $L = \log_\gamma (1-\gamma)$ and for each policy σ^l, let $\hat{s} = \operatorname{argmax}_s f(s, \sigma^l(s))$. Then for all $k > l + L$, it holds that $\sigma^k(\hat{s}) \neq \sigma^l(\hat{s})$.*

Finally, we aggregate the elimination steps to provide a polynomial upper bound on the total number of iterations. Since an action is eliminated from the set of potential maximizers every L steps, the algorithm must terminate within $O(n \cdot m \cdot L)$ iterations for $n = |\mathcal{S}|$ and $m = |\mathcal{A}|$. The full proof of Theorem 17 is available in Appendix N.

Theorem 17 *`RMDP-PI` terminates with an optimal policy in $\mathcal{O}\left(n \cdot m \cdot \frac{\log(1-\gamma)}{\log \gamma}\right)$ iterations.*

The policy evaluation step of `RMDP-PI` is strongly polynomial due to Corollary 10. The policy improvement is also trivially strongly polynomial, implying the following corollary immediately:

Corollary 18 *`RMDP-PI` is strongly polynomial given an RMDP $\mathcal{R}$ with n states, m actions and constant discount factor γ.*

6. Conclusion

In this paper, we studied discounted (s, a)-rectangular RMDPs with $\mathbf{L}_\infty$ uncertainty sets, a model that captures classical MDPs and turn-based stochastic games. Our main contribution is to resolve a fundamental algorithmic open problem for this setting. We showed that a robust policy iteration algorithm terminates in strongly polynomial time when the discount factor is fixed. Several directions remain open and relevant. While the present work establishes the theoretical foundations, it is an interesting future direction to complement the worst-case analysis with empirical studies of robust policy iteration on data-driven uncertainty sets, and to explore whether the potential-based arguments suggest improved practical variants or stopping criteria. Another related major question is whether policy iteration for MDPs admits polynomial or strongly polynomial bounds. Fearnley (2010) showed that one variant can take exponentially many iterations for discounted-sum MDPs, leaving open whether other variants run in polynomial time.

Acknowledgments

We thank Morteza Saghafian for his helpful suggestions when preparing the paper. This work was supported by Vienna Science and Technology Fund (WWTF), State of Lower Austria [Grant ID 10.47379/ICT25017], ERC CoG 863818 (ForM-SMArt) and Austrian Science Fund (FWF) 10.55776/COE12.

References

Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT Press, 2008.

Bahram Behzadian, Marek Petrik, and Chin Pang Ho. Fast algorithms for l_∞-constrained s-rectangular robust mdps. In *NeurIPS*, pages 25982–25992, 2021.

Krishnendu Chatterjee, Ehsan Kafshdar Goharshady, Mehrdad Karrabi, Petr Novotnỳ, and Đorđe Žikelić. Solving long-run average reward robust mdps via stochastic games. In *IJCAI*, pages 6707–6715. ijcai.org, 2024.

Karina Valdivia Delgado, Leliane N. de Barros, Daniel B. Dias, and Scott Sanner. Real-time dynamic programming for markov decision processes with imprecise probabilities. *Artif. Intell.*, 230:192–223, 2016.

Francois d'Epenoux. A probabilistic production and inventory problem. *Management Science*, 10 (1):98–108, 1963.

Cyrus Derman. *Finite State Markovian Decision Processes*. Academic Press, 1972.

John Fearnley. Exponential lower bounds for policy iteration. In *ICALP (2)*, volume 6199 of *Lecture Notes in Computer Science*, pages 551–562. Springer, 2010.

Tony Feng, Trieu H. Trinh, Garrett Bingham, Yuri Chervonyi, Junehyuk Jung, Dawsen Hwang, Joonkyung Lee, Carlo Pagano, Sang-hyun Kim, Federico Pasqualotto, Sergei Gukov, Junsu Kim, Kaiying Hou, Jonathan Lee, Golnaz Ghiasi, Yi Tay, YaGuang Li, Chenkai Kuang, Yuan Liu, Hanzhao (Maggie) Lin, Evan Liu, Nigamaa Nayakanti, Xiaomeng Yang, Heng-tze Cheng, Koray Kavukcuoglu, Quoc V. Le, and Minh-Thang Luong. Towards autonomous mathematics research. 2026.

Robert Givan, Sonia M. Leach, and Thomas L. Dean. Bounded-parameter markov decision processes. *Artif. Intell.*, 122(1-2):71–109, 2000.

Thomas Dueholm Hansen, Peter Bro Miltersen, and Uri Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16, 2013.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

David L. Kaufman and Andrew J. Schaefer. Robust modified policy iteration. *INFORMS J. Comput.*, 25(3):396–410, 2013.

Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994.

Marnix Suilen, Thom Badings, Eline M. Bovy, David Parker, and Nils Jansen. Robust markov decision processes: A place where AI and formal methods meet. In *Principles of Verification (3)*, volume 15262 of *Lecture Notes in Computer Science*, pages 126–154. Springer, 2024.

Yinyu Ye. The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. *Math. Oper. Res.*, 36(4):593–603, 2011.

Appendix A. Reduction from Turn Based Stochastic Games to L_∞ RMDPs

A turn based stochastic game G is defined as a tuple $(\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_r, c, \mathit{Succ}, P)$, where

- $\mathcal{S}_1, \mathcal{S}_2$ and $\mathcal{S}_r$ are disjoint sets. $\mathcal{S}_1$ is a set of player-1 states, $\mathcal{S}_2$ is a set of player-2 states, and $\mathcal{S}_r$ is a set of randomized states. Let $\mathcal{S} = \mathcal{S}_1 \uplus \mathcal{S}_2 \uplus \mathcal{S}_r$
- $c\colon \mathcal{S} \to \mathbb{R}$ is a cost function
- $\mathit{Succ}\colon \mathcal{S}_1 \cup \mathcal{S}_2 \to 2^{\mathcal{S}}$ specifies the set of possible successors from each $\mathcal{S}_1$ and $\mathcal{S}_2$ state.
- $P\colon \mathcal{S}_r \to \Delta(\mathcal{S})$ is the randomized transition function from $\mathcal{S}_r$ states.

The semantics are defined by player strategies as follows: a randomized strategy σ_i of player-i is a function $\sigma_i\colon \mathcal{S}_i \to \Delta(\mathcal{S})$ where $\mathit{support}(\pi_i(s)) \subseteq \mathit{Succ}(s)$. A run of G is then an infinite sequence $\pi_0, \pi_1 \ldots$ where if $\pi_j \in \mathcal{S}_i$ for some $i \in \{1, 2\}$, then $\Pr[\pi_{j+1}|\pi_j] = \sigma_i(\pi_j)[\pi_{j+1}]$, and if $\pi_j \in \mathcal{S}_r$, then $\Pr[\pi_{j+1}|\pi_j] = P(\pi_j)[\pi_{j+1}]$. The cylinder construction of Baier and Katoen (2008) synthesizes the full probability distribution over infinite sequences of states in G. For a discount factor γ, the discounted sum of costs over a run π is then defined as

$$c(\pi) = \sum_{j=0}^{\infty} \gamma^j c(\pi_j)$$

The goal of player-1 is to minimize the expected cost of the induced run while player-2 wants to maximize it. Hence the value of the game is defined as

$$\boldsymbol{v}_G(s) = \min_{\sigma_1} \max_{\sigma_2} \mathbb{E}_{\pi \sim (\sigma_1, \sigma_2)}[c(\pi)]$$

We show that any game $G = (\mathcal{S}_1^G, \mathcal{S}_2^G, \mathcal{S}_r^G, c^G, \mathit{Succ}^G, P^G)$ is equivalent to an RMDP $\mathcal{R} = (\mathcal{S}^{\mathcal{R}}, c^{\mathcal{R}}, \mathcal{A}^{\mathcal{R}}, \mathit{Succ}^{\mathcal{R}}, \mathcal{P}^{\mathcal{R}})$ where $\mathcal{P}^{\mathcal{R}}$ is an $\mathbf{L}_\infty$ uncertainty set specified by the nominal transition probabilities $\hat{P}^{\mathcal{R}}(s, a)$ and the uncertainty radii $\delta(s, a)$ for each state-action pair (s, a). The RMDP $\mathcal{R}$ is constructed as follows:

- $\mathcal{S}^{\mathcal{R}} = \mathcal{S}_1^G \cup \mathcal{S}_2^G \cup \mathcal{S}_r^G$
- $c^{\mathcal{R}} = c^G$
- For each $s \in \mathcal{S}_1^G$ and $s' \in \mathit{Succ}^G(s)$, there is an action $a \in \mathcal{A}^{\mathcal{R}}$ where $\mathit{Succ}^{\mathcal{R}}(s, a) = \{s'\}$, $\hat{P}^{\mathcal{R}}(s, a)[s'] = 1$ and $\delta(s, a) = 0$
- For each $s \in \mathcal{S}_r^G$, there is an action $a \in \mathcal{A}^{\mathcal{R}}$ where $\mathit{Succ}^{\mathcal{R}}(s, a) = \mathit{support}(P^G(s))$, $\hat{P}^{\mathcal{R}}(s, a)[s'] = P^G(s)[s']$ and $\delta^{\mathcal{R}}(s, a) = 0$.
- For each $s \in \mathcal{S}_2^G$, there is an action $a \in \mathcal{A}^{\mathcal{R}}$ where $\mathit{Succ}^{\mathcal{R}}(s, a) = \mathit{Succ}^G(s)$, $\hat{P}^{\mathcal{R}}(s, a)$ is an arbitrary distribution in $\Delta(\mathit{Succ}^G(s))$, and $\delta(s, a) = 1$.

We show that every strategy profile in G has an equivalent policy profile in $\mathcal{R}$ and vice versa.

1. Let (σ_1, σ_2) be a strategy profile in G. We construct a policy profile (σ, τ) in $\mathcal{R}$ that generates the same probability distribution over the set of infinite sequences of states in $\mathcal{S}^{\mathcal{R}}$ simply because $\Pr_{(\sigma_1,\sigma_2)}(s)[s'] = \Pr_{(\sigma,\tau)}[s](s')$. Specifically, $\sigma(s) = \sigma_1(s)$ for all $s \in \mathcal{S}_1 \cup \mathcal{S}_r$. Note that for $s \in \mathcal{S}_2$, $|\mathcal{A}^{\mathcal{R}}(s)| = 1$, hence $\sigma(s)$ is defined in the trivial manner. For τ, the environment has a single choice in all $s \in \mathcal{S}_1 \cup \mathcal{S}_r$. For $s \in \mathcal{S}_2$, we let $\tau(s) = \sigma_2(s)$. This is specifically a valid choice for the environment because $\mathcal{P}^{\mathcal{R}}(s) = \Delta(\mathit{Succ}(s))$. By construction, $\Pr_{(\sigma_1,\sigma_2)}(s)[s'] = \Pr_{(\sigma,\tau)}[s](s')$ for all states $s, s' \in \mathcal{S}^{\mathcal{R}}$.
2. Let (σ, τ) be a policy profile in $\mathcal{R}$. We construct (σ_1, σ_2) to be a strategy profile in G which induces the same transition function as (σ, τ). Given the construction of $\mathcal{R}$, it is possible to put $\sigma_1(s) = \sigma(s)$ for all $s \in \mathcal{S}_1$. For σ_2, let $s \in \mathcal{S}_2$ be arbitrary. We define $\sigma_2(s) = \tau(s, a)$ for the

unique action a defined at s in $\mathcal{A}^{\mathcal{R}}$. Again, by construction, $\Pr_{(\sigma_1,\sigma_2)}(s)[s'] = \Pr_{(\sigma,\tau)}[s](s')$ for all states $s, s' \in \mathcal{S}^{\mathcal{R}}$.

Appendix B. Proof of Lemma 1

Proof

- Fix $s \in \mathcal{S}$. Without loss of generality, we assume $(\mathcal{T}\boldsymbol{u})_s \geq (\mathcal{T}\boldsymbol{v})_s$. Let $\mathbf{p}'$ and $\mathbf{p}''$ denote the maximizers in $\mathcal{P}(s)$ for $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. We write the difference as follows:

$$\begin{aligned}
(\mathcal{T}\boldsymbol{u})_s - (\mathcal{T}\boldsymbol{v})_s &= \gamma(\mathbf{p}'\boldsymbol{u} - \mathbf{p}''\boldsymbol{v}) && \text{(Definition of } \mathcal{T}\text{)} \\
&\leq \gamma(\mathbf{p}'\boldsymbol{u} - \mathbf{p}'\boldsymbol{v}) && \text{(Optimality of } \mathbf{p}''\text{)} \\
&= \gamma\mathbf{p}'(\boldsymbol{u} - \boldsymbol{v}) && \text{(Linearity)} \\
&\leq \gamma \left\|\boldsymbol{u} - \boldsymbol{v}\right\|_\infty . && \text{(}\mathbf{p}'\text{ is a distribution)}
\end{aligned}$$

  Since this bound holds for an arbitrary state s, the result follows.
- For any state $s \in \mathcal{S}$, we have:

$$\begin{aligned}
(\mathcal{T}\boldsymbol{v}^\tau)_s &= \mathbf{c}_s + \gamma \max_{\mathbf{p}\in\mathcal{P}(s)} \mathbf{p}^\top \boldsymbol{v}^\tau && \text{(Definition of } \mathcal{T}\text{)} \\
&\geq \mathbf{c}_s + \gamma\mathcal{P}_s^\tau \cdot \boldsymbol{v}^\tau && \text{(Since } \mathcal{P}_s^\tau \in \mathcal{P}(s)\text{)} \\
&= \boldsymbol{v}^\tau && \text{Definition of } \boldsymbol{v}^\tau
\end{aligned}$$

  Since this inequality holds for all states, the result follows.
- Direct result of the Banach fixed-point theorem.

■

Appendix C. Proof of Lemma 2

Proof

- Let $\mathcal{P}^t$ denote the transition matrix corresponding to the policy in iteration t. We analyze the difference between the value vectors:

$$\begin{aligned}
\boldsymbol{v}^{t+1} - \boldsymbol{v}^t &= (\mathbb{I} - \gamma\mathcal{P}^{t+1})^{-1}\mathbf{c} - \boldsymbol{v}^t && \text{(By definition of value)} \\
&= (\mathbb{I} - \gamma\mathcal{P}^{t+1})^{-1}(\mathbb{I} - \gamma\mathcal{P}^{t+1})\left[(\mathbb{I} - \gamma\mathcal{P}^{t+1})^{-1}\mathbf{c} - \boldsymbol{v}^t\right] && \text{(Multiplying by } \mathbb{I}\text{)} \\
&= (\mathbb{I} - \gamma\mathcal{P}^{t+1})^{-1}\left[\mathbf{c} - (\mathbb{I} - \gamma\mathcal{P}^{t+1})\boldsymbol{v}^t\right] && \text{(Rearranging)} \\
&= (\mathbb{I} - \gamma\mathcal{P}^{t+1})^{-1}\left[(\mathbf{c} + \gamma\mathcal{P}^{t+1}\boldsymbol{v}^t) - \boldsymbol{v}^t\right] && \text{(Expanding inner terms)} \\
&= (\mathbb{I} - \gamma\mathcal{P}^{t+1})^{-1}\left[\mathcal{T}\boldsymbol{v}^t - \boldsymbol{v}^t\right] && \text{(Definition of } \mathcal{T} \text{ and } \tau^{t+1}\text{)}
\end{aligned}$$

  By Lemma 1, $\mathcal{T}\boldsymbol{v}^t - \boldsymbol{v}^t \succcurlyeq \mathbf{0}$. We express the inverse matrix using the Neumann series:

$$(\mathbb{I} - \gamma\mathcal{P}^{t+1})^{-1} = \sum_{i=0}^{\infty}(\gamma\mathcal{P}^{t+1})^i.$$

  Since γ and all entries of $\mathcal{P}^{t+1}$ are non-negative, the sum consists of non-negative terms. Consequently, $(\mathbb{I} - \gamma\mathcal{P}^{t+1})^{-1}$ is component-wise non-negative. This implies $\boldsymbol{v}^{t+1} - \boldsymbol{v}^t \succcurlyeq \mathbf{0}$.

- First note that it can be easily shown that $\boldsymbol{v}^{t+1} \succcurlyeq \mathcal{T}\boldsymbol{v}^t$ as follows:

$$\begin{aligned}
\boldsymbol{v}^{t+1} - \mathcal{T}\boldsymbol{v}^t &= \boldsymbol{v}^{t+1} - (\mathbf{c} + \gamma\mathcal{P}^{t+1}\boldsymbol{v}^t) && \text{(Definition of } \mathcal{T} \text{ and } \tau^{t+1}\text{)} \\
&= \boldsymbol{v}^{t+1} - \big((\mathbb{I} - \gamma\mathcal{P}^{t+1})\boldsymbol{v}^{t+1} + \gamma\mathcal{P}^{t+1}\boldsymbol{v}^t\big) && \text{(Replacing } \mathbf{c} \text{ using Equation 1)} \\
&= \boldsymbol{v}^{t+1} - (\boldsymbol{v}^{t+1} - \gamma\mathcal{P}^{t+1}\boldsymbol{v}^{t+1} + \gamma\mathcal{P}^{t+1}\boldsymbol{v}^t) && \text{(Expanding)} \\
&= \gamma\mathcal{P}^{t+1}(\boldsymbol{v}^{t+1} - \boldsymbol{v}^t) && \text{(Simplifying)}
\end{aligned}$$

Since $\gamma > 0$, $\mathcal{P}^{t+1} \geq 0$ entrywise, and by the first bullet point $\boldsymbol{v}^{t+1} - \boldsymbol{v}^t \succcurlyeq \mathbf{0}$, we conclude $\boldsymbol{v}^{t+1} \succcurlyeq \mathcal{T}\boldsymbol{v}^t$.

We now proceed by proving the second statement of the Lemma:

We use induction on t. The base case $t = 0$ is trivial. For $t > 0$:

$$\begin{aligned}
\left\|\boldsymbol{v}^t - \boldsymbol{v}^*\right\|_\infty &\leq \left\|\mathcal{T}\boldsymbol{v}^{t-1} - \boldsymbol{v}^*\right\|_\infty && \text{(Since } \boldsymbol{v}^* \succcurlyeq \boldsymbol{v}^t \succcurlyeq \mathcal{T}\boldsymbol{v}^{t-1} \text{ as above)} \\
&= \left\|\mathcal{T}\boldsymbol{v}^{t-1} - \mathcal{T}\boldsymbol{v}^*\right\|_\infty && \text{(Fixed point property by Lemma 1)} \\
&\leq \gamma\left\|\boldsymbol{v}^{t-1} - \boldsymbol{v}^*\right\|_\infty && \text{(Contraction of } \mathcal{T} \text{ by Lemma 1)} \\
&\leq \gamma^t\left\|\boldsymbol{v}^0 - \boldsymbol{v}^*\right\|_\infty && \text{(By induction)}
\end{aligned}$$

■

Appendix D. Proof of Lemma 4

Proof

$$\begin{aligned}
\boldsymbol{v}_s^* - \boldsymbol{v}_s^\tau &= (\mathbf{c} + \gamma\mathcal{P}^*\boldsymbol{v}^*)_s - (\mathbf{c} + \gamma\mathcal{P}^\tau\boldsymbol{v}^\tau)_s \\
&\geq (\mathbf{c} + \gamma\mathcal{P}^*\boldsymbol{v}^*)_s - (\mathbf{c} + \gamma\mathcal{P}^\tau\boldsymbol{v}^*)_s && \text{(Since } \boldsymbol{v}^* \succcurlyeq \boldsymbol{v}^\tau\text{)} \\
&= \gamma\left(\mathcal{P}_s^* - \mathcal{P}_s^\tau\right)\boldsymbol{v}^* && \text{(Simplifying)}
\end{aligned}$$

Let $\epsilon := \min(\mathcal{P}^*_{s,s'} - \mathcal{P}^\tau_{s,s'}, \mathcal{P}^\tau_{s,s''} - \mathcal{P}^*_{s,s''})$. By the symmetry of f under s', s'' and the fact that $\epsilon \leq 0$ cannot produce the argmax, we can assume without loss of generality that $\epsilon \geq 0$. Now if $f_\tau(s, s', s'') \leq 0$ the bound is trivial since $\boldsymbol{v}_s^* \geq \boldsymbol{v}_s^\tau$. Now, consider perturbing $\mathcal{P}_s^\tau$ by increasing $\mathcal{P}^\tau_{s,s'}$ by ϵ and decreasing $\mathcal{P}^\tau_{s,s''}$ by ϵ. Let $\mathcal{P}_s^{\tau'}$ denote the resulting vector. By construction $0 \leq \epsilon \leq \mathcal{P}^*_{s,s'} - \mathcal{P}^\tau_{s,s'}$, so $\mathcal{P}^\tau_{s,s'} + \epsilon \leq \mathcal{P}^*_{s,s'} \leq \hat{P}_{s,s'} + \delta_s$. Similarly $\mathcal{P}^\tau_{s,s''} - \epsilon \geq \mathcal{P}^*_{s,s''} \geq \hat{P}_{s,s''} - \delta_s$. The row sum is preserved. The perturbation satisfies:

$$\mathcal{P}_s^{\tau'}\boldsymbol{v}^* - \mathcal{P}_s^\tau\boldsymbol{v}^* = \epsilon\left(\boldsymbol{v}^*_{s'} - \boldsymbol{v}^*_{s''}\right) = f_\tau(s, s', s'').$$

We now derive the bound using the optimality of $\mathcal{P}^*$:

$$\begin{aligned}
& \mathcal{P}_s^*\boldsymbol{v}^* \geq \mathcal{P}_s^{\tau'}\boldsymbol{v}^* && \text{(Optimality of } \mathcal{P}^*\text{)} \\
\implies\quad & \mathcal{P}_s^*\boldsymbol{v}^* - \mathcal{P}_s^\tau\boldsymbol{v}^* \geq \mathcal{P}_s^{\tau'}\boldsymbol{v}^* - \mathcal{P}_s^\tau\boldsymbol{v}^* && \text{(Subtracting } \mathcal{P}_s^\tau\boldsymbol{v}^*\text{)} \\
\implies\quad & \left(\mathcal{P}_s^* - \mathcal{P}_s^\tau\right)\boldsymbol{v}^* \geq \epsilon\left(\boldsymbol{v}^*_{s'} - \boldsymbol{v}^*_{s''}\right) && \text{(Substitution)} \\
\implies\quad & \left(\mathcal{P}_s^* - \mathcal{P}_s^\tau\right)\boldsymbol{v}^* \geq f_\tau(s, s', s'') && \text{(Definition of } f_\tau\text{)}
\end{aligned}$$

Combining this result with the first inequality in the proof, we obtain:

$$\boldsymbol{v}_s^* - \boldsymbol{v}_s^\tau \geq \gamma f_\tau(s, s', s'').$$

■

Appendix E. Proof of Lemma 5

Proof First, we relate the value difference to the transition difference using the Bellman equation.

$$\begin{aligned}
\boldsymbol{v}^* - \boldsymbol{v}^\tau &= (\mathbb{I} - \gamma\mathcal{P}^*)^{-1}\mathbf{c} - (\mathbb{I} - \gamma\mathcal{P}^\tau)^{-1}\mathbf{c} && \text{(By Equation 1)} \\
&= (\mathbb{I} - \gamma\mathcal{P}^\tau)^{-1}\left[(\mathbb{I} - \gamma\mathcal{P}^\tau) - (\mathbb{I} - \gamma\mathcal{P}^*)\right](\mathbb{I} - \gamma\mathcal{P}^*)^{-1}\mathbf{c} && \text{(Using } A^{-1} - B^{-1} = B^{-1}(B-A)A^{-1}\text{)} \\
&= (\mathbb{I} - \gamma\mathcal{P}^\tau)^{-1}\left[\gamma(\mathcal{P}^* - \mathcal{P}^\tau)\right]\boldsymbol{v}^* && \text{(Simplifying terms)}
\end{aligned}$$

We now take the infinity norm of both sides:

$$\begin{aligned}
\|\boldsymbol{v}^* - \boldsymbol{v}^\tau\|_\infty &= \gamma\left\|(\mathbb{I} - \gamma\mathcal{P}^\tau)^{-1}(\mathcal{P}^* - \mathcal{P}^\tau)\boldsymbol{v}^*\right\|_\infty \\
&\leq \gamma\left\|(\mathbb{I} - \gamma\mathcal{P}^\tau)^{-1}\right\|_\infty\|(\mathcal{P}^* - \mathcal{P}^\tau)\boldsymbol{v}^*\|_\infty \quad \text{(Submultiplicativity of } \mathbf{L}_\infty \text{ norm)}
\end{aligned}$$

We analyze the term $\left\|(\mathbb{I} - \gamma\mathcal{P}^\tau)^{-1}\right\|_\infty$ using the Neumann series. Since $\mathcal{P}^\tau$ is a transition matrix, it is row-stochastic (all rows sum to 1 and entries are non-negative). Consequently, its induced infinity norm is $\|\mathcal{P}^\tau\|_\infty = 1$. Furthermore, the product of stochastic matrices is stochastic, implying $\left\|(\mathcal{P}^\tau)^t\right\|_\infty = 1$ for all $t \geq 0$.

$$\begin{aligned}
\left\|(\mathbb{I} - \gamma\mathcal{P}^\tau)^{-1}\right\|_\infty &= \left\|\sum_{t=0}^{\infty}(\gamma\mathcal{P}^\tau)^t\right\|_\infty && \text{(Neumann Series expansion)} \\
&\leq \sum_{t=0}^{\infty}\gamma^t\left\|(\mathcal{P}^\tau)^t\right\|_\infty && \text{(Triangle inequality)} \\
&= \sum_{t=0}^{\infty}\gamma^t \cdot 1 && \text{(Since } (\mathcal{P}^\tau)^t \text{ is stochastic)} \\
&= \frac{1}{1-\gamma} && \text{(Geometric series sum)}
\end{aligned}$$

Substituting this back into the previous inequality:

$$\|\boldsymbol{v}^* - \boldsymbol{v}^\tau\|_\infty \leq \frac{\gamma}{1-\gamma}\|(\mathcal{P}^* - \mathcal{P}^\tau)\boldsymbol{v}^*\|_\infty .$$

It remains to bound the term $\|(\mathcal{P}^* - \mathcal{P}^\tau)\boldsymbol{v}^*\|_\infty$. We do this by interpreting the difference matrix as a collection of mass transfers.

Mass transfer interpretation. We express the difference $\mathcal{P}^* - \mathcal{P}^\tau$ as a sequence of flows between coordinates. Define a tensor $m \in \mathbb{R}^{n\times n\times n}$ where $m(i,j,k)$ represents the probability mass moved to state j (where τ has a deficit) from state k (where τ has excess) in row i. Since both $\mathcal{P}^\tau$ and $\mathcal{P}^*$ lie within the same convex $\mathbf{L}_\infty$-ball, the vector connecting them ($\mathcal{P}^* - \mathcal{P}^\tau$) can be decomposed into a finite sum of *feasible* elementary redistributions. For $i \in \mathcal{S}$ Suppose $K_i \subseteq \mathcal{S}\times\mathcal{S}$ is such that $\mathcal{P}_i^\tau + \sum_{(j,k)\in K_i} m(i,j,k)(e_j - e_k) = \mathcal{P}_i^*$, where e_j is the j-th standard basis vector. We may further assume the decomposition satisfies $\boldsymbol{v}_j^* \geq \boldsymbol{v}_k^*$ whenever $m(i,j,k) > 0$: otherwise, by the optimality of $\mathcal{P}_i^*$, swapping such a pair back would yield a feasible distribution with strictly larger dot product against $\boldsymbol{v}^*$. This also gives $(\mathcal{P}_i^* - \mathcal{P}_i^\tau)\boldsymbol{v}^* \geq 0$ in every row, so the absolute values below can be dropped.

Bounding the total change. We analyze the row i that maximizes the difference:

$$\begin{aligned}\|(\mathcal{P}^* - \mathcal{P}^\tau)\boldsymbol{v}^*\|_\infty &= \max_i (\mathcal{P}_i^* - \mathcal{P}_i^\tau)\boldsymbol{v}^* && \text{(Optimality of } \mathcal{P}^*\text{)}\\ &= \max_i \sum_{(j,k)\in K_i} m(i,j,k)(\boldsymbol{v}_j^* - \boldsymbol{v}_k^*) && \text{(Decomposing into transfers)}\end{aligned}$$

Note that for any specific triplet (i,j,k), the contribution to the sum $m(i,j,k)\left|\boldsymbol{v}_j^* - \boldsymbol{v}_k^*\right|$ is bounded by the potential function $f_\tau(i,j,k)$ which is upper-bounded by $f_\tau(s,s',s'')$:

$$\begin{aligned}\max_i \sum_{(j,k)\in K_i} m(i,j,k)(\boldsymbol{v}_j^* - \boldsymbol{v}_k^*) &\le \max_i \sum_{(j,k)\in K_i} f_\tau(i,j,k) && \text{(Definition of } f_\tau\text{)}\\ &\le \sum_{(j,k)\in K_i} f_\tau(s,s',s'') && \text{(Bounding by global max)}\\ &\le n^2 f_\tau(s,s',s'') && \text{(Summing over at most } n^2 \text{ pairs)}\end{aligned}$$

Substituting this back into our earlier bound:

$$\|\boldsymbol{v}^* - \boldsymbol{v}^\tau\|_\infty \le \frac{\gamma}{1-\gamma}\left(n^2 f_\tau(s,s',s'')\right).$$

This completes the proof. ■

Appendix F. Proof of Lemma 6

Proof We chain the upper bound of Lemma 5 (for τ) and the lower bound of Lemma 4 (for τ') to derive the result. First, we apply the upper bound to the policy τ:

$$\begin{aligned}\|\boldsymbol{v}^* - \boldsymbol{v}^\tau\|_\infty &\le \frac{n^2\gamma}{1-\gamma} f_\tau(s,s',s'') && \text{(By Lemma 5)}\\ &\le \frac{n^2\gamma}{1-\gamma}\cdot 2 f_{\tau'}(s,s',s'') && \text{(By definition of } f_{\tau'} \text{ and assumption)}\end{aligned}$$

The last inequality is true because $\boldsymbol{v}_{s'}^* \ge \boldsymbol{v}_{s''}^*$ since we are considering the maximum potential. Finally, we link this to the value error of τ' using the lower bound:

$$\|\boldsymbol{v}^* - \boldsymbol{v}^\tau\|_\infty \le \frac{2n^2}{1-\gamma}\left(\gamma f_{\tau'}(s,s',s'')\right) \le \frac{2n^2}{1-\gamma}\left(\boldsymbol{v}_s^* - \boldsymbol{v}_s^{\tau'}\right) \le \frac{2n^2}{1-\gamma}\left\|\boldsymbol{v}^* - \boldsymbol{v}^{\tau'}\right\|_\infty$$

Where the second inequality comes from Lemma 4 and the third one is due to the definition of $\mathbf{L}_\infty$ norm. Rearranging the terms gives the claimed bound: $\left\|\boldsymbol{v}^* - \boldsymbol{v}^{\tau'}\right\|_\infty \ge \frac{1-\gamma}{2n^2}\|\boldsymbol{v}^* - \boldsymbol{v}^\tau\|_\infty$. ■

Appendix G. Proof of Lemma 7

Proof Let s, s', s'' be the critical triplet defined in the lemma. Assume, for the sake of contradiction, that there exists some step $l > t + L$ such that:

$$\min\left(\mathcal{P}^*_{s,s'} - \mathcal{P}^{\tau^l}_{s,s'},\ \mathcal{P}^{\tau^l}_{s,s''} - \mathcal{P}^*_{s,s''}\right) > \frac{1}{2}\min\left(\mathcal{P}^*_{s,s'} - \mathcal{P}^{\tau^t}_{s,s'},\ \mathcal{P}^{\tau^t}_{s,s''} - \mathcal{P}^*_{s,s''}\right).$$

This assumption allows us to invoke Lemma 6, providing a lower bound on the error at step l. Simultaneously, standard Policy Iteration convergence provides an upper bound. We combine these as follows:

$$\left\|\boldsymbol{v}^{\tau^l} - \boldsymbol{v}^*\right\|_\infty \leq \gamma^{l-t}\left\|\boldsymbol{v}^{\tau^t} - \boldsymbol{v}^*\right\|_\infty \qquad \text{(By Lemma 2)}$$

$$\left\|\boldsymbol{v}^{\tau^l} - \boldsymbol{v}^*\right\|_\infty \geq \frac{1-\gamma}{2n^2}\left\|\boldsymbol{v}^{\tau^t} - \boldsymbol{v}^*\right\|_\infty \qquad \text{(By Lemma 6 and assumption)}$$

Combining these two inequalities implies:

$$\implies \quad \gamma^{l-t} \geq \frac{1-\gamma}{2n^2}$$

$$\implies \quad l - t \leq \log_\gamma\left(\frac{1-\gamma}{2n^2}\right) \qquad \text{(Taking } \log_\gamma\text{, inequality flips since } \gamma < 1\text{).}$$

This conclusion ($l - t \leq L$) directly contradicts the premise that $l > t + L$. Thus, the assumption is false, and the difference must have decreased by at least half. ■

Appendix H. Proof of Lemma 8

Proof This result follows immediately from Theorem 20. This theorem bounds the number of distinct scales (powers of 2) that can be generated by linear combinations with small integer coefficients. ■

Let us begin with the definition.

Definition 19 *We call a* dyadic interval *any interval either of the form* $[2^i, 2^{i+1})$ *or of the form* $[-2^{i+1}, -2^i)$ *for some integer* i. $[0, 0]$ *is treated as a separate dyadic interval.*

Observe that the collection of dyadic intervals offers a countable partition of $\mathbb{R}$. For $i \in \mathbb{Z}$, we say that a real number has degree i if it belongs to the component $[2^i, 2^{i+1})$ or to the component $[-2^{i+1}, -2^i)$.

For a finite subset $X \subseteq \mathbb{R}$ of the real numbers, and a positive real C we denote by $A(X, C)$ the set of real numbers of the form

$$\sum_{x \in X} f(x) \cdot x,$$

where $f : X \to \mathbb{Z}$ is any function with $\|f\|_\infty \leq C$. For a subset Y of $\mathbb{R}$, we denote by

$$\mathrm{Dyad}(Y),$$

the collection of dyadic intervals containing elements of Y.

Theorem 20 *Let C be in $\mathbb{R}_{>0}$. Then for all X finite subset of $\mathbb{R}$ we have that*

$$|\mathrm{Dyad}(A(X,C))| = O_C(|X| \cdot \log(|X|)).$$

Shortly after and independently, an effective strengthening of Theorem 20 was discovered and established by a custom mathematics research agent, named *Aletheia*, built upon Gemini Deep Think at Google DeepMind under the lead of Tony Feng, see Feng et al. (2026).

Aletheia's theorem goes as follows.

Theorem 21 *Let C be in $\mathbb{R}_{>0}$. Then for all X finite subset of $\mathbb{R}$ we have that*

$$|\mathrm{Dyad}(A(X,C))| \leq 2 \cdot (2|X| - 1)(\lfloor \log_2(2|X|C + 1) \rfloor + 2) + 1.$$

The prompt that the last author of the present paper fed to *Aletheia* was as follows:
Let C be a positive integer. For X a finite set of real numbers, denote with $A(X,C)$ the set of real numbers obtained by taking integer linear combinations of elements of X, with coefficients bounded by C. Find an upper bound, which is polynomial in $|X|$, for the number of intervals of the form $[2^n, 2^{n+1})$, with n nonnegative integer, that intersect an element of $A(X,C)$. If possible find an asymptotically sharp upper bound.

At the center of both arguments to prove Theorem 20 and Theorem 21, there is an application of Siegel's Lemma. However this is implemented quite differently in the two proofs.

Overview of proof of Theorem 20 Our idea to prove Theorem 20 is inspired by a standard move in modern number theory, which is to examine a function field analogue first. In this case the dyadic subdivision corresponds simply to the degree partition of the set of polynomials [2]. The analogue for function fields of Theorem 20 is the following simple proposition that we shall prove right away. The proof will provide us with a clear road map for the proof of Theorem 20.

Proposition 22 *Let $\mathbb{F}$ be a finite field and fix m a nonnegative integer. Let X be a subset of the polynomial ring $\mathbb{F}[T]$. Then one sees at most $(m+1) \cdot |X|$ degrees as one runs in the set of non-zero polynomials of the form*

$$\sum_{p \in X} f(p) \cdot p,$$

where $\deg(f(p)) \leq m$ *for each p in X.*

Proof *The set described is the $\mathbb{F}$-vector space V generated by the elements in $X \cdot \{1, \ldots, t^m\}$. Since it is $\mathbb{F}$-spanned by $(m+1) \cdot |X|$ elements, this vector space V has $\mathbb{F}$-dimension at most $(m+1) \cdot |X|$. However, $(m+1) \cdot |X| + 1$ non-zero polynomials of pairwise distinct degree always form, in particular, a linearly independent set. But then V cannot contain a linearly independent subset exceeding its dimension. Thus, there are no such $(m+1) \cdot |X| + 1$ non-zero polynomials to be seen in V, yielding the desired conclusion.* ■

With the proof of Proposition 22 in our hands, we can now briefly overview the main ideas of the proof of Theorem 20. To mimic the proof, we proceed as follows:
(1) In the context of Theorem 20, the field $\mathbb{F}$ is replaced by $[-C_1, C_1] \cap \mathbb{Z}$, which does not have

2. We shall not make the analogy entirely strict by making X to be a subset of the completion of the function field with respect to the degree valuation, as the purpose of the function field analogy is to simply illustrate the proof in an idealized setting.

any standard algebraic structure. Nevertheless, we can mimic the notion of linear independence of N real numbers: we say that N reals are "C_1-linearly independent" in case they don't have a non-trivial linear dependency using only integers of absolute value going up to C_1.
(2) We now observe that for any positive real C_1, there exists a positive real D such that if we have any collection of non-zero real numbers whose degrees are pairwise spaced by D, then the collection is C_1-linearly independent. This is a very simple fact, and it is Proposition 24 below.
(3) On the other hand, by construction, our N reals in $A(X, C)$ are the image of an integer matrix of size at most C. Siegel's lemma from the geometry of numbers shows that they must satisfy a non-trivial linear dependency of size about (at most) $(CN)^{\frac{n}{N-n}}$, where $n := |X|$. For N of the size $n \cdot \log(n)$, this bound on the size of the linear dependency is no more than $C \cdot C_0$, as n runs over all positive integers, for C_0 some universal positive constant.
(4) Let us now take D as in part (2) relative to $C_1 := C \cdot C_0$. If now we witness (say) at least $100 \cdot D \cdot |X| \cdot \log(|X|)$ dyadic intervals in $A(X, C)$, then by a greedy search, we can find about at least $|X| \cdot \log(|X|)$ non-zero values in $A(X, C)$ pairwise spaced by about D. Now by step (1) they are $C \cdot C_0$-linearly independent. On the other hand, in (3) we have witnessed a non-zero $C \cdot C_0$-linear dependency between these numbers, which is a contradiction.

Overview of the proof of Theorem 21 The following overview is directly extracted from *Aletheia*'s output, which gives a remarkably clear high level view of the steps. Here is *Aletheia*'s summary:
Method Sketch: Let $k = |X|$. We aim to bound $N(X, C)$, the number of intervals $[2^n, 2^{n+1})$ ($n \geq 0$) intersecting $A(X, C)$. We focus on the set $Y = A(X, C) \cap [1, \infty)$.
The upper bound is derived using a Gap Principle established via Siegel's Lemma. We define $R = 2kC + 1$. We prove that any sequence in Y where consecutive terms have a ratio of at least R (an R-separated sequence) has length at most $2k - 1$. This is shown by contradiction: assuming a sequence of length $2k$, Siegel's Lemma implies a non-trivial integer linear relation among the elements with coefficients bounded by $2kC$. However, the R-separation property with $R > 2kC$ forbids such a relation.
We then cover Y by a union of at most $2k - 1$ intervals, each of the form $[s, Rs)$. The number of dyadic intervals intersecting such an interval is bounded by $\lfloor \log_2 R \rfloor + 2$. This yields the upper bound $N(X, C) \leq (2k - 1)(\lfloor \log_2(2kC + 1) \rfloor + 2)$. This bound is $O(k(\log k + \log C))$.

We remark that Siegel's lemma is used in a different regime in the two proofs of Theorem 20 and Theorem 21, respectively. In the former it directly controls combinations of uniformly bounded size, while in the latter it allows combinations of size $|X|$.

Proof of Theorem 20 Let us formalize the notion of linear independence explained above.

Definition 23 *Let N be a positive integer and C_1 a positive real number. We say that a collection $y_1, \ldots, y_N$ of real numbers is C_1-linearly independent in case*

$$(y_1, \ldots, y_N) \cdot \underline{x} \neq 0,$$

for all non-zero vectors $\underline{x}$ in $\mathbb{Z}^N$ such that

$$\|\underline{x}\|_\infty \leq C_1.$$

They are said C_1-linearly dependent if they are not C_1-linearly independent.

The following fact can be viewed as a criterion for "linear independence" of a set of real numbers over a set of coefficients with bounded size: it is the analogue of the observation that over $\mathbb{F}[T]$ a set of non-zero polynomials of different degrees must be linearly independent.

Proposition 24 *Let C_1 be a positive real. Then there exists a positive integer D such that the following holds. Let $N \geq 2$ be an integer. Let $(y_1, \ldots, y_N)$ in $\mathbb{R}^N$ such that $y_1 \neq 0$ and for each $1 \leq i < N$, we have that $\deg(y_{i+1}) > D + \deg(y_i)$.*

Then we have that $y_1, \ldots, y_N$ is C_1-linear independent.

Proof *Let $\underline{x}$ be in $\mathbb{Z}^N$ such that*

$$(y_1, \ldots, y_N) \cdot \underline{x} = 0.$$

It will suffice to show that $x_N = 0$, and then proceed by induction on N, which will give that $x_i = 0$ for each $i \geq 2$. At that point, it must follow that $x_1 = 0$, since $y_1 \neq 0$ by assumption.

To this end, notice that, as soon as D is sufficiently large, we have the preliminary bound

$$|y_{i+1}| > 2 \cdot |y_i|,$$

for all $1 \leq i < N$. This implies in particular that

$$|y_{N-1}| > \sum_{1 \leq i \leq N-2} |y_i|.$$

On the other hand, when D is sufficiently large, we also have the bound

$$|y_N| > 2 \cdot C_1 \cdot |y_{N-1}|.$$

Thus, we deduce that, if x_N is non-zero, since it is an integer, it must be that $|x_N \cdot y_N| \geq |y_N|$ and therefore

$$|x_N||y_N| \geq |y_N| > 2C_1 \cdot |y_{N-1}| \geq C_1 \cdot (|y_{N-1}| + \ldots + |y_1|) \geq |\sum_{1 \leq i \leq N-1} x_i y_i|,$$

which directly contradicts

$$(y_1, \ldots, y_N) \cdot \underline{x} = 0.$$

This contradiction stemmed from the assumption that $x_N \neq 0$, and hence it must be that $x_N = 0$. Now we can proceed inductively on the rest of the coefficients and find that they are all 0, where in the base case we use that $y_1 \neq 0$ by assumption, thus yielding that $\underline{x} = 0$. ■

We need the following application of Siegel's lemma.

Proposition 25 *There exists an absolute constant $C_0 > 0$ such that the following holds. For all positive integers n and any real number C at least 1, and any matrix A with coefficients in $\mathbb{Z}$ with $\|A\|_\infty \leq C$, n columns and N rows, with $n \cdot \log(N) \geq N \geq n \cdot \log(n) - 1$, one has a non-zero vector $\underline{x} \in \mathbb{Z}^N$ in the left-kernel of A with*

$$\|\underline{x}\|_\infty \leq C_0 \cdot C.$$

Proof *Siegel's lemma tells us that there is a vector $\underline{x}$ of norm at most*

$$(\|A\|_\infty N)^{\frac{n}{N-n}} \leq (C \cdot (n \cdot log(n)))^{\frac{1}{log(n)-1-\frac{1}{n}}} \leq C \cdot (n \cdot log(n))^{\frac{1}{log(n)-1-\frac{1}{n}}}.$$

The last quantity in the product converges to e as n goes to ∞, in particular, it stays bounded, yielding a C_0 as desired. ■

We are now in a position to prove the following.

Proof [Proof of Theorem 20] Let C_0 be as in Proposition 25. Take now any choice of D coming from Proposition 24, where we place $C_1 := C_0 \cdot C$. Suppose now that there is a non-empty set X of positive reals with $|\text{Dyad}(A(X,C))| \geq 100 \cdot D \cdot |X| \cdot \log(|X|)$. That means, by a greedy search, that we can arrange to have a positive integer $|X| \cdot \log(|X|) \geq N \geq |X| \cdot \log(|X|) - 1$ and a vector

$$(y_1, \ldots, y_N) \in A(X,C)^N,$$

such that $y_1 \neq 0$ and for each $1 \leq i < N$, we have that $\deg(y_{i+1}) > D + \deg(y_i)$. In virtue of our choice of D and of Proposition 24, we see that $\underline{y}$ cannot be orthogonal to any non-zero integer vector of $\|\cdot\|_\infty$-norm at most $C_0 \cdot C$. On the other hand, by construction, $\underline{y}$ is in the image of an integer matrix A having $|X|$ columns, $|X|\log(|X|) \geq N \geq |X|\log(|X|) - 1$ rows and $\|A\|_\infty \leq C$. Hence, the element provided by Proposition 25 in the left kernel is a non-zero integer vector of norm at most $C \cdot C_0$, which is orthogonal to anything in the image of A, since the kernel of the transpose equals the orthogonal to the image. This gives a contradiction. Therefore, it must be that

$$|\text{Dyad}(A(X,C))| \leq 100 \cdot D \cdot |X| \cdot \log(|X|),$$

which is the desired conclusion. ■

Proof of Theorem 21 *Aletheia* was prompted to search for elements in $A(X,C) \cap [1,\infty)$. Clearly whatever upper bound B one finds for the number of dyadic intervals touched by $A(X,C) \cap [1,\infty)$, then provides a bound of the form $2B+1$ for $\text{Dyad}(A(X,C))$.

The text that follows is directly extracted from the output of *Aletheia*.

We utilize Siegel's Lemma to establish a Gap Principle.

Lemma 26 (Siegel's Lemma) *Let B be an $M \times N$ matrix with integer entries, with $N > M > 0$. Let $H \geq 1$ be an upper bound for the absolute values of the entries of B. Then there exists a non-zero integer vector $c \in \mathbb{Z}^N \setminus \{0\}$ such that $Bc = 0$ and*

$$\|c\|_\infty \leq (NH)^{M/(N-M)}.$$

Proof See, for example, M. Hindry and J. H. Silverman, Diophantine Geometry: An Introduction, Springer, 2000, Part D, Lemma 4.1. ■

A sequence of positive real numbers (z_i) is called R-separated if $z_{i+1}/z_i \geq R$ for all i.

Lemma 27 (Gap Principle) *Let $k \geq 1, C \geq 1$. Let $R = 2kC + 1$. Any R-separated sequence of distinct positive elements in $A(X,C)$ has length $m \leq 2k - 1$.*

Proof Let $C' = 2kC$. Note that $R = C' + 1$. Suppose for contradiction that there exists an R-separated sequence $0 < z_1 < z_2 < \cdots < z_m$ in $A(X,C)$ with $m = 2k$.

For each $i = 1, \ldots, m$, $z_i \in A(X,C)$, so $z_i = \sum_{j=1}^k a_{ij} x_j$ with $a_{ij} \in \mathbb{Z}$ and $|a_{ij}| \leq C$. Let $v_i = (a_{i1}, \ldots, a_{ik})$.

Let B be the $k \times m$ matrix whose columns are the vectors v_i. The entries of B are bounded by $H = C$. We apply Lemma 26 with $M = k$ and $N = m = 2k$. Since $N > M$, there exists a non-zero integer vector $c = (c_1, \ldots, c_m) \in \mathbb{Z}^m$ such that $Bc = 0$. The bound on the coefficients is

$$\|c\|_\infty \leq (NH)^{M/(N-M)} = (2kC)^{k/(2k-k)} = 2kC = C'.$$

The condition $Bc = 0$ implies a linear relation among the elements z_i:

$$\sum_{i=1}^{m} c_i z_i = \sum_{i=1}^{m} c_i \sum_{j=1}^{k} a_{ij} x_j = \sum_{j=1}^{k} x_j \left(\sum_{i=1}^{m} c_i a_{ij} \right) = 0.$$

Let J be the largest index such that $c_J \neq 0$. Then $c_J z_J = -\sum_{i=1}^{J-1} c_i z_i$. Taking absolute values (and noting $z_i > 0$):

$$|c_J| z_J = \left| \sum_{i=1}^{J-1} c_i z_i \right| \leq \sum_{i=1}^{J-1} |c_i| z_i \leq C' \sum_{i=1}^{J-1} z_i.$$

Since the sequence (z_i) is R-separated and increasing, $z_i \leq R^{-(J-1-i)} z_{J-1}$ for $i \leq J-1$.

$$\sum_{i=1}^{J-1} z_i \leq z_{J-1} \sum_{l=0}^{J-2} R^{-l} < z_{J-1} \sum_{l=0}^{\infty} R^{-l} = z_{J-1} \frac{R}{R-1}.$$

Also, by the separation property, $z_J \geq R z_{J-1}$. Combining these inequalities:

$$|c_J| R z_{J-1} \leq |c_J| z_J < C' z_{J-1} \frac{R}{R-1}.$$

Since $z_{J-1} > 0$ and $R > 0$, we can divide by $R z_{J-1}$:

$$|c_J| < \frac{C'}{R-1}.$$

We chose $R = C' + 1$, so $R - 1 = C'$.

$$|c_J| < \frac{C'}{C'} = 1.$$

This contradicts the fact that c_J is a non-zero integer. Therefore, the length of the sequence must be $m \leq 2k - 1$. ■

Lemma 28 *Let $I = [a, b)$ be an interval with $1 \leq a < b$. The number of dyadic intervals $[2^n, 2^{n+1})$ with $n \geq 0$ that intersect I, denoted by $D(I)$, is bounded by*

$$D(I) \leq \lfloor \log_2(b/a) \rfloor + 2.$$

Proof Since $a \geq 1$, the exponents n must be non-negative. The smallest exponent n_{min} is $\lfloor \log_2 a \rfloor$. The largest exponent n_{max} must satisfy $2^{n_{max}} < b$, so $n_{max} = \lceil \log_2 b \rceil - 1$. The count is $D(I) = n_{max} - n_{min} + 1 = \lceil \log_2 b \rceil - \lfloor \log_2 a \rfloor$. Let $R = b/a$. Let $r = \log_2 R$ and $x = \log_2 a$. $D(I) = \lceil r + x \rceil - \lfloor x \rfloor$. Let $I_r = \lfloor r \rfloor$, $f_r = \{r\}$, and $I_x = \lfloor x \rfloor$, $f_x = \{x\}$. $D(I) = \lceil I_r + f_r + I_x + f_x \rceil - I_x = I_r + \lceil f_r + f_x \rceil$. Since $0 \leq f_r, f_x < 1$, we have $0 \leq f_r + f_x < 2$. Thus $\lceil f_r + f_x \rceil \in \{0, 1, 2\}$. Therefore, $D(I) \leq I_r + 2 = \lfloor \log_2 R \rfloor + 2$. ■

We now prove the upper bound in Theorem 21.

Proof [Proof of the Upper Bound] Let $Y = A(X, C) \cap [1, \infty)$. If Y is empty, $N(X, C) = 0$. Assume Y is non-empty. Y is finite. Let $R = 2kC + 1$. We construct a maximal R-separated subsequence of Y greedily. Let $s_1 = \min Y$. Since $Y \subset [1, \infty)$, $s_1 \geq 1$. For $j \geq 1$, if s_j is defined, let $Y_j = \{y \in Y : y \geq Rs_j\}$. If Y_j is non-empty, define $s_{j+1} = \min Y_j$. Otherwise, the sequence terminates. Let the sequence be $S = \{s_1, \dots, s_m\}$. By construction, $s_{j+1} \geq Rs_j$. By Lemma 27, $m \leq 2k - 1$.

We show that Y is covered by the union of the intervals $J_j = [s_j, Rs_j)$ for $j = 1, \dots, m$. Let $y \in Y$. Since $s_1 = \min Y$, $y \geq s_1$. Let J be the largest index such that $s_J \leq y$. If $J = m$. The sequence terminated because Y_m is empty. Thus, any $y' \in Y$ with $y' \geq s_m$ must satisfy $y' < Rs_m$. So $y \in J_m$. If $J < m$. Then s_{J+1} exists, and $y < s_{J+1}$ by the maximality of J. If we had $y \geq Rs_J$, then $y \in Y_J$. By definition, $s_{J+1} = \min Y_J \leq y$, a contradiction. Thus $y < Rs_J$. So $y \in J_J$. Therefore, $Y \subset \bigcup_{j=1}^m J_j$.

The total number of dyadic intervals intersecting Y is bounded by the sum of the counts for each J_j. By Lemma 28, since $s_j \geq 1$ and the ratio for J_j is R, the number of dyadic intervals intersecting J_j is at most $\lfloor \log_2 R \rfloor + 2$.

$$N(X, C) \leq \sum_{j=1}^{m} D(J_j) \leq m(\lfloor \log_2 R \rfloor + 2) \leq (2k-1)(\lfloor \log_2(2kC+1) \rfloor + 2).$$

■

This upper bound is $O(k \log(kC)) = O(k(\log k + \log C))$.

Appendix I. Proof of Theorem 9

Proof Let $\tau^0, \tau^1, \dots, \tau^T$ be the sequence of policies generated by the algorithm.
1. Finite Discrepancy Set. For a state s with nominal transition vector $\hat{P}_s$ and uncertainty radius δ_s, the realized transition probability $P = \mathcal{P}_s^\tau$ in any policy τ results from redistributing mass under the $\mathbf{L}_\infty$ constraint, according to Algorithm 2. Given the properties of such policies, discussed in Section 3, we define X_s as follows so that $\mathcal{P}_{s,s'}^\tau \in A(X_s)$ for all $s, s' \in \mathcal{S}$:

$$X_s = \{\hat{P}_{s,s'} \mid s' \in \mathcal{S}\} \cup \{1\} \cup \{\delta(s), 2 \cdot \delta(s), \cdots, n \cdot \delta(s)\}.$$

Given the properties of homotopic policies generated by Algorithm 2, it can be verified that all possible values of $\mathcal{P}_{s,s'}^* - \mathcal{P}_{s,s'}^\tau$ and $\mathcal{P}_{s,s'}^\tau - \mathcal{P}_{s,s'}^*$ are included in $A(X_s)$ when we set the constant $c = 2$. Note that $|X_s| = 2n + 1$. Lemma 8 implies that $A(X_s)$ has polynomially many different MSBs:

$$\mathrm{Deg}(A(X_s)) \in \mathcal{O}(|X_s| \log |X_s|) = \mathcal{O}(n \log n).$$

So, $\mathcal{P}_{s,s'}^* - \mathcal{P}_{s,s'}^\tau$ and $\mathcal{P}_{s,s'}^\tau - \mathcal{P}_{s,s'}^*$ also can take polynomially many MSBs.
2. Convergence Rate. Let

$$f_t(s, s', s'') = \min\left(\mathcal{P}_{s,s'}^* - \mathcal{P}_{s,s'}^{\tau^t},\ \mathcal{P}_{s,s''}^{\tau^t} - \mathcal{P}_{s,s''}^*\right)(\boldsymbol{v}_{s'}^* - \boldsymbol{v}_{s''}^*)$$

be the potential value at iteration t. Let $(s_t, s_t', s_t'') = \mathrm{argmax}_{s,s',s''} f_{\tau^t}(s, s', s'')$ be the maximizing triple. Lemma 7 states that for $L = \log_\gamma\left(\frac{1-\gamma}{2n^2}\right)$, the mass transfer quantity associated with this triple decreases by a factor of at least 2 after L steps. In terms of logarithmic degree, the most significant bit of this quantity decreases by at least 1 every L steps.

3. Total Complexity. There are at most n^3 distinct triples (s, s', s''). For any specific triple, the mass transfer value can take at most $\mathcal{O}(n \log n)$ different logarithmic scales (from Step 1) before it vanishes or the triple is no longer active. Since each scale reduction requires at most L steps, we bound the total number of steps by:

$$\text{Total Iterations} \leq n^3 \times \mathcal{O}(n \log n) \times L = \mathcal{O}(n^4 \log n \cdot L).$$

We substitute the definition of L and omit constant factors to obtain:

$$\mathcal{O}\left(n^4 \log n \cdot L\right) = \mathcal{O}\left(n^4 \log n \cdot \frac{\log\left(\frac{1-\gamma}{n^2}\right)}{\log \gamma}\right).$$

■

Appendix J. Proof of Lemma 11

Proof

- To prove contraction, fix $s \in \mathcal{S}$. Let $a^* = \operatorname{argmin}_{a\in\mathcal{A}} \max_{\mathbf{p}\in\mathcal{P}_{s,a}} \mathbf{p}^\top \boldsymbol{v}$ and $\mathbf{p}' = \operatorname{argmax}_{\mathbf{p}\in\mathcal{P}_{s,a^*}} \mathbf{p}^\top \boldsymbol{u}$. Then:

$$\begin{aligned}
(\mathcal{T}\boldsymbol{u})_s - (\mathcal{T}\boldsymbol{v})_s &= \gamma\left(\min_{a\in\mathcal{A}} \max_{\mathbf{p}\in\mathcal{P}_{s,a}} \mathbf{p}^\top \boldsymbol{u} - \min_{a\in\mathcal{A}} \max_{\mathbf{p}\in\mathcal{P}_{s,a}} \mathbf{p}^\top \boldsymbol{v}\right) && \text{(By Definition)}\\
&\leq \gamma\left(\max_{\mathbf{p}\in\mathcal{P}_{s,a^*}} \mathbf{p}^\top \boldsymbol{u} - \max_{\mathbf{p}\in\mathcal{P}_{s,a^*}} \mathbf{p}^\top \boldsymbol{v}\right) && \text{(By definition of } a^*\text{)}\\
&\leq \gamma \mathbf{p}'^\top(\boldsymbol{u} - \boldsymbol{v}) && \text{(By definition of } \mathbf{p}'\text{)}\\
&\leq \gamma \left\|\boldsymbol{u} - \boldsymbol{v}\right\|_\infty && \text{(Since } \mathbf{p}' \text{ is a distribution)}
\end{aligned}$$

  Symmetry yields the absolute value bound.

-

$$\begin{aligned}
(\mathcal{T}\boldsymbol{v}^\sigma)_s &= \mathbf{c}_s + \gamma \min_{a\in\mathcal{A}} \max_{p\in\mathcal{P}_{s,a}} p^\top \boldsymbol{v}^\sigma && \text{(Definition of } \mathcal{T}\text{)}\\
&\leq \mathbf{c}_s + \gamma \max_{p\in\mathcal{P}_{s,\sigma(s)}} p^\top \boldsymbol{v}^\sigma && \text{(Restricting min to } \sigma(s)\text{)}\\
&= \boldsymbol{v}^\sigma_s && \text{(Definition of } \boldsymbol{v}^\sigma\text{)}
\end{aligned}$$

- The fixed-point property follows directly from the Banach fixed-point theorem.

■

Appendix K. Proof of Lemma 12

Proof

- $$\begin{aligned}
\boldsymbol{v}^t - \boldsymbol{v}^{t+1} &= \boldsymbol{v}^t - (\mathbb{I} - \gamma P^{t+1})^{-1}\mathbf{c} && \text{(By Equation 2)} \\
&= (\mathbb{I} - \gamma P^{t+1})^{-1}(\mathbb{I} - \gamma P^{t+1})\left[\boldsymbol{v}^t - (\mathbb{I} - \gamma P^{t+1})^{-1}\mathbf{c}\right] && \text{(Multiplying by } \mathbb{I}\text{)} \\
&= (\mathbb{I} - \gamma P^{t+1})^{-1}\left[(\mathbb{I} - \gamma P^{t+1})\boldsymbol{v}^t - \mathbf{c}\right] && \text{(Expanding)} \\
&= (\mathbb{I} - \gamma P^{t+1})^{-1}\left[\boldsymbol{v}^t - (\mathbf{c} + \gamma P^{t+1}\boldsymbol{v}^t)\right] && \text{(Rearranging terms)} \\
&\succcurlyeq (\mathbb{I} - \gamma P^{t+1})^{-1}\left[\boldsymbol{v}^t - \mathcal{T}\boldsymbol{v}^t\right] && \text{(By Definition of } \mathcal{T}\text{)}
\end{aligned}$$

  By Lemma 11, $\boldsymbol{v}^t - \mathcal{T}\boldsymbol{v}^t \succcurlyeq \mathbf{0}$. Since $(\mathbb{I} - \gamma P^{t+1})^{-1} \succcurlyeq \mathbf{0}$, the result follows.
- We first observe that $\mathbf{c} + \gamma P^{t+1}\boldsymbol{v}^t \preccurlyeq \mathcal{T}\boldsymbol{v}^t$: for each state s, since $P_s^{t+1} \in \mathcal{P}(s, \sigma^{t+1}(s))$ is one feasible point,

  $$P_s^{t+1}\boldsymbol{v}^t \ \leq \max_{\mathbf{p}\in\mathcal{P}(s,\,\sigma^{t+1}(s))} \mathbf{p}^\top \boldsymbol{v}^t \ = \ \min_{a\in\mathcal{A}} \max_{\mathbf{p}\in\mathcal{P}(s,a)} \mathbf{p}^\top \boldsymbol{v}^t,$$

  where the equality uses that $\sigma^{t+1}(s)$ is the greedy improvement w.r.t. $\boldsymbol{v}^t$. Multiplying by γ and adding $\mathbf{c}_s$ yields the claim.
  We now show $\boldsymbol{v}^{t+1} \preccurlyeq \mathcal{T}\boldsymbol{v}^t$ as follows:

  $$\begin{aligned}
\mathcal{T}\boldsymbol{v}^t - \boldsymbol{v}^{t+1} &\succcurlyeq \mathbf{c} + \gamma P^{t+1}\boldsymbol{v}^t - \boldsymbol{v}^{t+1} && (\mathbf{c} + \gamma P^{t+1}\boldsymbol{v}^t \preccurlyeq \mathcal{T}\boldsymbol{v}^t) \\
&= (\mathbb{I} - \gamma P^{t+1})\boldsymbol{v}^{t+1} + \gamma P^{t+1}\boldsymbol{v}^t - \boldsymbol{v}^{t+1} && \text{(Substituting } \mathbf{c}\text{)} \\
&= \boldsymbol{v}^{t+1} - \gamma P^{t+1}\boldsymbol{v}^{t+1} + \gamma P^{t+1}\boldsymbol{v}^t - \boldsymbol{v}^{t+1} && \text{(Expanding)} \\
&= \gamma P^{t+1}(\boldsymbol{v}^t - \boldsymbol{v}^{t+1}) && \text{(Simplifying)}
\end{aligned}$$

  Since $\gamma > 0$, $P^{t+1} \geq 0$, and $\boldsymbol{v}^t \succcurlyeq \boldsymbol{v}^{t+1}$ (proved above), we conclude $\boldsymbol{v}^{t+1} \preccurlyeq \mathcal{T}\boldsymbol{v}^t$.
  Now, to prove the second statement of the lemma we use induction on t. The base case $t = 0$ holds trivially. For $t > 0$:

  $$\begin{aligned}
\left\|\boldsymbol{v}^t - \boldsymbol{v}^*\right\|_\infty &\leq \left\|\mathcal{T}\boldsymbol{v}^{t-1} - \boldsymbol{v}^*\right\|_\infty && (\boldsymbol{v}^* \preccurlyeq \boldsymbol{v}^t \preccurlyeq \mathcal{T}\boldsymbol{v}^{t-1}) \\
&= \left\|\mathcal{T}\boldsymbol{v}^{t-1} - \mathcal{T}\boldsymbol{v}^*\right\|_\infty && \text{(Since } \boldsymbol{v}^* \text{ is the fixed point)} \\
&\leq \gamma\left\|\boldsymbol{v}^{t-1} - \boldsymbol{v}^*\right\|_\infty && \text{(By contraction in Lemma 11)} \\
&\leq \gamma^t\left\|\boldsymbol{v}^0 - \boldsymbol{v}^*\right\|_\infty && \text{(By induction)}
\end{aligned}$$

■

Appendix L. Proof of Lemma 14

Proof

- **(Lower-bound)**

$$\begin{aligned}
\boldsymbol{v}_s^\sigma - \boldsymbol{v}_s^* &= \left(\mathbf{c}_s + \gamma \max_{\mathbf{p}\in\mathcal{P}_{s,a}} \mathbf{p}^\top \boldsymbol{v}^\sigma\right) - \left(\mathbf{c}_s + \gamma \max_{\mathbf{p}\in\mathcal{P}_{s,\sigma^*(s)}} \mathbf{p}^\top \boldsymbol{v}^*\right) && \text{(By Equation 2)} \\
&= \gamma \left(\max_{\mathbf{p}\in\mathcal{P}_{s,a}} \mathbf{p}^\top \boldsymbol{v}^\sigma - \max_{\mathbf{p}\in\mathcal{P}_{s,\sigma^*(s)}} \mathbf{p}^\top \boldsymbol{v}^*\right) && \text{(Simplification)} \\
&\geq \gamma \left(\max_{\mathbf{p}\in\mathcal{P}_{s,a}} \mathbf{p}^\top \boldsymbol{v}^* - \max_{\mathbf{p}\in\mathcal{P}_{s,\sigma^*(s)}} \mathbf{p}^\top \boldsymbol{v}^*\right) && \text{(Since } \boldsymbol{v}^\sigma \succcurlyeq \boldsymbol{v}^*\text{)} \\
&= \gamma f(s,a) && \text{(By Definition)}
\end{aligned}$$

- **(Upper-bound)** Assume τ is environment's best response to σ and τ^* is environment's best response to σ^*. Denote the transition matrix when fixing policies by $P^{\sigma,\tau}$ and P^*. We begin by expanding the value difference using the value definition from Equation 2:

$$\begin{aligned}
\boldsymbol{v}^\sigma - \boldsymbol{v}^* &= (\mathbb{I} - \gamma P^{\sigma,\tau})^{-1}\mathbf{c} - (\mathbb{I} - \gamma P^*)^{-1}\mathbf{c} \\
&= \Big((\mathbb{I} - \gamma P^{\sigma,\tau})^{-1} - (\mathbb{I} - \gamma P^*)^{-1}\Big)\mathbf{c}
\end{aligned}$$

We apply the matrix identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ to the term inside the norm:

$$\begin{aligned}
\boldsymbol{v}^\sigma - \boldsymbol{v}^* &= (\mathbb{I} - \gamma P^{\sigma,\tau})^{-1}\Big((\mathbb{I} - \gamma P^*) - (\mathbb{I} - \gamma P^{\sigma,\tau})\Big)(\mathbb{I} - \gamma P^*)^{-1}\mathbf{c} \\
&= \gamma\, (\mathbb{I} - \gamma P^{\sigma,\tau})^{-1}(P^{\sigma,\tau} - P^*)(\mathbb{I} - \gamma P^*)^{-1}\mathbf{c}
\end{aligned}$$

Replacing $(\mathbb{I} - \gamma P^*)^{-1}\mathbf{c}$ by $\boldsymbol{v}^*$ (based on Equation 2) we have the following equation:

$$\boldsymbol{v}^\sigma - \boldsymbol{v}^* \;=\; \gamma(\mathbb{I} - \gamma P^{\sigma,\tau})^{-1}(P^{\sigma,\tau} - P^*)\boldsymbol{v}^*.$$

We first bound the vector $(P^{\sigma,\tau} - P^*)\boldsymbol{v}^*$ component-wise. Fix $s \in \mathcal{S}$. Since τ is the environment's best response to σ, we have $P_s^{\sigma,\tau} \in \mathcal{P}_{s,\sigma(s)}$, so:

$$\begin{aligned}
\big((P^{\sigma,\tau} - P^*)\boldsymbol{v}^*\big)_s &= P_s^{\sigma,\tau}\boldsymbol{v}^* - P_s^*\boldsymbol{v}^* && \text{(Definition of } P_s^{\sigma,\tau} \text{ and } P_s^*\text{)} \\
&\leq \max_{\mathbf{p}\in\mathcal{P}_{s,\sigma(s)}} \mathbf{p}^\top \boldsymbol{v}^* - \max_{\mathbf{p}\in\mathcal{P}_{s,\sigma^*(s)}} \mathbf{p}^\top \boldsymbol{v}^* && \text{(Taking the max over } \mathcal{P}_{s,\sigma(s)}\text{)} \\
&= f(s,\sigma(s)). && \text{(Definition of } f\text{)}
\end{aligned}$$

Define the vector $\mathbf{f}^\sigma \in \mathbb{R}^n$ by $\mathbf{f}_s^\sigma := f(s,\sigma(s))$. The componentwise bound above gives

$$(P^{\sigma,\tau} - P^*)\boldsymbol{v}^* \;\preccurlyeq\; \mathbf{f}^\sigma.$$

Since $P^{\sigma,\tau}$ is stochastic, the matrix $(\mathbb{I} - \gamma P^{\sigma,\tau})^{-1} = \sum_{i\geq 0}(\gamma P^{\sigma,\tau})^i$ has only nonnegative entries, so multiplying both sides preserves the inequality. Combined with $\boldsymbol{v}^\sigma \succcurlyeq \boldsymbol{v}^*$, we obtain the nonnegative bounding chain

$$\mathbf{0} \;\preccurlyeq\; \boldsymbol{v}^\sigma - \boldsymbol{v}^* \;\preccurlyeq\; \gamma(\mathbb{I} - \gamma P^{\sigma,\tau})^{-1}\mathbf{f}^\sigma.$$

Taking $\|\cdot\|_\infty$ preserves inequalities between nonnegative vectors, so:

$$\begin{aligned}
\|\boldsymbol{v}^\sigma - \boldsymbol{v}^*\|_\infty &\le \gamma \left\|(\mathbb{I} - \gamma P^{\sigma,\tau})^{-1}\mathbf{f}^\sigma\right\|_\infty && \text{(Bounding chain above)}\\
&\le \gamma \left\|(\mathbb{I} - \gamma P^{\sigma,\tau})^{-1}\right\|_\infty \cdot \|\mathbf{f}^\sigma\|_\infty && \text{(Submultiplicativity of } \mathbf{L}_\infty \text{ norm)}\\
&\le \frac{\gamma}{1-\gamma} \cdot \|\mathbf{f}^\sigma\|_\infty && \text{(Neumann series; } P^{\sigma,\tau} \text{ stochastic)}\\
&= \frac{\gamma}{1-\gamma} \cdot \max_{s \in \mathcal{S}} f(s, \sigma(s)) && \text{(Definition of } \mathbf{L}_\infty \text{ norm and } \mathbf{f}^\sigma)\\
&= \frac{\gamma}{1-\gamma} f(\hat{s}, \sigma(\hat{s})). && \text{(Definition of } \hat{s})
\end{aligned}$$

This completes the proof.

■

Appendix M. Proof of Lemma 16

Proof Assume for the sake of contradiction that there exists $k > l + L$ such that $\sigma^k(\hat{s}) = \sigma^l(\hat{s})$. Since the value vectors are non-increasing, $\boldsymbol{v}^{\sigma^l} \ge \boldsymbol{v}^{\sigma^k}$. Applying Lemma 15:

$$\left\|\boldsymbol{v}^{\sigma^k} - \boldsymbol{v}^*\right\|_\infty \ge (1-\gamma) \left\|\boldsymbol{v}^{\sigma^l} - \boldsymbol{v}^*\right\|_\infty .$$

However, from the global convergence rate (Lemma 12), we know:

$$\left\|\boldsymbol{v}^{\sigma^k} - \boldsymbol{v}^*\right\|_\infty \le \gamma^{k-l} \left\|\boldsymbol{v}^{\sigma^l} - \boldsymbol{v}^*\right\|_\infty .$$

We note that if $\|\mathbf{v}^{\sigma^l} - \mathbf{v}^*\|_\infty = 0$, then σ^l is optimal, and the algorithm terminates at iteration l. Since the sequence continues to $k > l$, we must have $\|\mathbf{v}^{\sigma^l} - \mathbf{v}^*\|_\infty > 0$. We can therefore divide both inequalities by this strictly positive term to obtain:

$$\log_\gamma(1-\gamma) \ge k - l \implies L \ge k - l.$$

This contradicts the assumption that $k > l + L$. ■

Appendix N. Proof of Theorem 17

Proof For each iteration ℓ, let $\hat{s}_\ell := \operatorname{argmax}_{s \in \mathcal{S}} f(s, \sigma^\ell(s))$ and call $C_\ell := (\hat{s}_\ell, \sigma^\ell(\hat{s}_\ell))$ the *critical pair* at ℓ. We show that each pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ can be critical only inside a window of length at most L.

Suppose a fixed pair (s, a) is critical at two iterations $\ell_1 < \ell_2$, i.e., $C_{\ell_1} = C_{\ell_2} = (s, a)$. Then $\hat{s}_{\ell_1} = s$ and $\sigma^{\ell_1}(s) = \sigma^{\ell_2}(s) = a$. Applying Lemma 16 at ℓ_1 (so that $\hat{s}$ in the lemma equals s) gives $\sigma^k(s) \ne a$ for all $k > \ell_1 + L$. Hence $\ell_2 \le \ell_1 + L$.

Since every iteration has a critical pair and there are at most $n \cdot m$ distinct pairs, each contributing a window of at most L iterations, the total number of iterations T is bounded by:

$$T \le n \cdot m \cdot L = n \cdot m \cdot \log_\gamma(1-\gamma) = n \cdot m \cdot \frac{\log(1-\gamma)}{\log \gamma}.$$

■