Revisiting Value Iteration: Unified Analysis of Discounted and Average-Reward Cases

Arsenii Mustafin
Cyber-physical Systems Group
Aalto University
Espoo, Finland 02150
arsenii.mustafin@aalto.fi

Xinyi Sheng
Cyber-physical Systems Group
Aalto University
Espoo, Finland 02150
xinyi.sheng@aalto.fi

Dominik Baumann
Cyber-physical Systems Group
Aalto University
Espoo, Finland 02150
dominik.baumann@aalto.fi

Abstract

While Value Iteration (VI) is one of the most fundamental algorithms in Reinforcement Learning, its theoretical convergence guarantees still exhibit a persistent mismatch with empirical behavior. In the discounted-reward case, classical theory guarantees geometric convergence with rate γ, while in the average-reward case recent work suggests that only sublinear convergence can be expected. In practice, however, VI is often observed to converge significantly faster. In this work, we show through a unified geometry-based analysis that, under an assumption of a unique and unichain optimal policy, *(i)* convergence is geometric in *both* the discounted- and average-reward settings and *(ii)* the convergence rate is faster than previous analyses suggest.

1 Introduction

Value Iteration (VI), introduced by Bellman in the late 1950s, is one of the oldest and most fundamental algorithms for solving Markov Decision Processes (MDPs). However, as we argue in this paper, its convergence properties are still not fully understood. Howard (1960) showed that, in discounted reward MDPs, VI converges geometrically in the ℓ^∞ norm with rate equal to the discount factor γ, that this rate is tight in the worst case, and that, as γ approaches 1, worst-case bounds become sublinear. The understanding of VI's convergence is still largely based on this early result. Moreover, Lee & Ryu (2025) suggest that in the average-reward setting with $\gamma = 1$, sublinear convergence is not only a worst-case bound, but also optimal. Thus, the case appears to be closed. Nevertheless, in this paper, we show that this impression is incomplete and potentially misleading.

Let us consider an instructive example, inspired by examples commonly used in RL theory lectures (e.g., by Szepesvári (2021)). We have an MDP with three groups of states, "heaven," "purgatory," and "hell." Each group itself consists of three states. Intuitively, heaven is the goal, forming a rewarding cycle under the optimal policy, with small probabilities of slipping into purgatory or hell. The full specification is given in Appendix A.

Suppose we use VI to solve this MDP. Given the discussion above, we should expect geometric convergence with rate γ in the discounted reward case and sublinear convergence as γ approaches 1. Nevertheless, when running VI on this MDP and plotting the span seminorm of the error vector

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

(normalized by its initial value) for different γ, Figure 1(A) shows that, even as γ approaches 1, the convergence rate does not become sublinear. In fact, we still see geometric convergence with a rate strictly smaller than 1. This result might, of course, be an artifact of a carefully handcrafted MDP. Thus, Figure 1(B) shows the convergence behavior of VI on a random MDP, which exhibits the same characteristics. Existing literature cannot explain these empirical results.

In this paper, we fill this gap through a unified analysis of the discounted and average-reward settings, which are typically treated separately, yielding new convergence rates for both. This has direct practical relevance. For instance, in modern reinforcement learning (RL), VI underlies the critic update in actor–critic methods, often implemented with neural networks as function approximators. When practitioners observe slow convergence, it is unclear whether this stems from approximation error, optimization issues, or the fundamental convergence behavior of VI itself. Sharp theoretical guarantees, as we provide in this paper, enable us to disentangle these sources of slowness.

Contributions. In this paper, we provide a new analysis of the VI algorithm convergence that unifies the discounted and average-reward settings. Our main contributions are as follows:

- We show that, in the discounted-reward case, under the assumption that the unique optimal policy is unichain, VI exhibits geometric convergence with rate $\iota^{1/n^2}\gamma < \gamma$ in the span seminorm. Thus, the total number of iterations required to obtain an ϵ-optimal policy is

$$\mathcal{O}\left(\frac{\log(1/\epsilon) + \log(1/(1-\gamma))}{\log(1/\gamma) + \frac{\log(1/\iota)}{n^2}}\right).$$

- We show that in the average-reward case, under the same assumption, the VI algorithm exhibits geometric convergence with a rate ι in terms of the span seminorm, which implies that the total number of iterations required to obtain an ϵ-optimal policy is

$$\mathcal{O}\left(\frac{\log(1/\epsilon)}{\frac{\log(1/\iota)}{n^2}}\right).$$

- We expand a geometric interpretation of MDPs (Mustafin et al., 2025) previously introduced for the discounted reward case onto the average-reward case. We show that in both cases, VI has the same dynamics, allowing us to analyze both cases together.

Previous Work. MDPs were introduced in the late 1950s, together with foundational algorithms such as VI (Bellman, 1957) and a comprehensive dynamic programming framework (Howard, 1960). Two standard performance criteria emerged: the discounted reward criterion, extensively summarized by Puterman (1994), and the average-reward criterion, developed through the works of Blackwell (1962), Yushkevich (1974), Denardo & Fox (1968), and characterized further by Schweitzer & Federgruen (1978); Schweitzer (1984).

For discounted MDPs, Howard (1960) showed that the convergence rate of VI is upper-bounded by the discount factor γ, and that this bound is attainable. Subsequent works refined and systematized these results and their assumptions (Puterman, 1994; Feinberg & Huang, 2014; Feinberg & He, 2020), establishing VI as a central tool for computing value functions in the discounted case. More recently Lee & Ryu (2023) show a sublinear convergence with a rate $1/t$.

For average-reward MDPs, convergence of VI has been extensively studied. In the unichain case, linear convergence rates can be obtained via the δ- and ergodicity coefficients (e.g., Seneta, 2006; Hübner & Hinderer, 1977; Puterman, 2014), and J-stage span contractions yield linear rates every J iterations in the span seminorm (Federgruen & Schweitzer, 1978; van der Wal, 1981; Puterman, 2014). In the multichain case, normalized VI iterates converge to the optimal average reward, although the policy error may not vanish (Puterman, 2014, Thm. 9.4.1). Necessary and sufficient conditions for VI convergence, as well as asymptotic linear rates on the Bellman error, were established by Schweitzer & Federgruen (1977, 1979). To ensure convergence of iterates to the bias function $h^\star$, White (1963) introduced Relative Value Iteration (RVI), with sufficient conditions studied by Morton & Wecker (1977), and additional conditions for VI convergence were studied by Vecchia et al. (2012). We go beyond those results by showing geometric convergence of VI also in the average-reward setting with respect to the span seminorm.

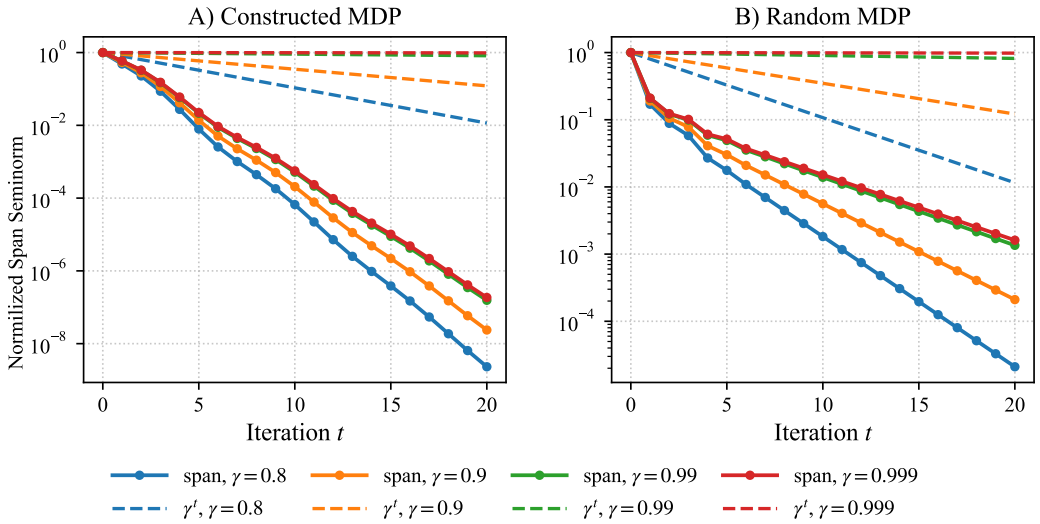


Figure 1: Normalized span seminorm versus iteration t (log scale) for four discount factors $\gamma \in \{0.8, 0.9, 0.99, 0.999\}$. Subfigure (A) shows an MDP used as an example in Section 1; a detailed description is given in Appendix A. Subfigure (B) shows a random MDP with 10 states, 1 to 4 actions on each state, and random rewards. Solid curves are the empirical spans (each normalized by its initial value). Across both MDPs, as γ approaches 1, the empirical spans decay rate approaches a geometrical convergence with a rate significantly faster than 1.

Several works have also investigated the relationship between discounted MDPs and average-reward MDPs. In particular, Puterman (1994) reveals a direct connection between discounted and average-reward quantities. Connections at the algorithmic level were studied by Tsitsiklis & Van Roy (2002). From an optimality perspective, discounted, average, and Blackwell optimality were unified through sensitive discount optimality (Mahadevan, 1996b), while Grand-Clément & Petrik (2023) showed that average optimality can be reduced to discounted optimality. These results demonstrate deep connections between the discounted- and average-reward settings, yet researchers typically treat the two cases separately through different value representations.

2 Preliminaries and Problem Statement

MDP Model. We employ a standard discrete finite MDP framework (Puterman, 2014), modifying it slightly by redefining the SAP space for convenience. That is, we define the MDP as the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S} = \{s_1, \ldots, s_n\}$ is a finite set of n states, and $\mathcal{A}$ is the set of all possible actions. Following Mustafin et al. (2025), we make the set of attainable state–action pairs the main object of our analysis and define it as $\mathcal{Z}$; we will use this concept extensively and, for convenience, refer to a state–action as a **SAP**. Since every SAP $z = (s, a)$ is tied to exactly one state, we define an auxiliary mapping st such that $\mathrm{st}(z) = s$ means that the state of SAP z is s ($z = (s, a')$ for some action a'). Each SAP z is characterized by a transition probability vector $\mathcal{P}(z) = (p_1^z, \ldots, p_n^z)$, where p_i^z defines the probability of transitioning to state s_i after selecting z, drawn from the set of all transition probabilities $\mathcal{P}$, and a deterministic reward $r^z \in [0, 1]$ drawn from the set of all rewards $\mathcal{R}$ (assuming scaled rewards). Lastly, $\gamma \in (0, 1]$ is a discount factor that specifies how much less a reward is valued if it is received one step later.

Given the set of SAPs, a choice rule (or **policy**) is a map $\pi : \mathcal{S} \to \mathcal{A}$. We consider only deterministic stationary policies, which means that for each state s, the same action is chosen whenever the state is visited. Therefore, it is convenient to view a **policy as a set of SAPs**, $\pi = \{z_1, \ldots, z_n\}$ with $z_i = (s_i, \pi(s_i))$. If a policy π chooses SAP z, we write $z \in \pi$.

Once the policy is fixed, the MDP dynamics reduce to a Markov Chain. A chain is called **unichain** if it contains a single recurrent class, possibly with a non-empty set of transient states. Otherwise, the MDP is called **multichain** following Puterman (2014).

For a given MDP $\mathcal{M}$ and policy π, performance can be measured in two ways corresponding to the discounted and average-reward criteria. We start with the former, as it is slightly simpler.

Discounted MDP. In the discounted reward case, the value of the policy π in state s, denoted as $V^\pi(s)$, is defined as the expected discounted reward obtained by starting from s and following π,$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | s\right]$, where r_t is the reward obtained at time t. The vector of values V^π is the unique vector satisfying the Bellman equation $V^\pi(s) = r^{\pi(s)} + \gamma \sum_{s'} p_{s'}^{\pi(s)} V^\pi(s')$. We can also write this in matrix form as $V^\pi = R^\pi + \gamma P^\pi V^\pi$, where R^π is a vector of all rewards of policy π, so that the ith entry $R^\pi(i) = r^{\pi(i)}$; and P^π is a transition kernel of the Markov chain induced by policy π. It is a square matrix whose ith row is equal to $P^\pi(i) = \mathcal{P}(\pi(i)) = (p_1^{\pi(i)}, \ldots, p_n^{\pi(i)})$.

In the discounted reward case, the optimal policy π^* is a policy that attains the maximum values in every state: $V^{\pi^*}(s) \geq V^\pi(s)$ for any π and s. We are also often interested in an approximate solution, or an ϵ-optimal policy π^ϵ ($\epsilon > 0$) such that $V^{\pi^*}(s) - V^{\pi^\epsilon}(s) < \epsilon$.

Average-reward MDP. In the average-reward case, we set $\gamma = 1$, i.e., all rewards are equally important to the agent, independent of when they are received. This case is considered more challenging, as tackling infinite values involves two quantities per policy: the *gain* and the *bias*.

For a given policy π, the *average reward* (or *gain*) from state s is basically equivalent to the values in the discounted case with $\gamma = 1$, but we average along time, expressed as $\rho^\pi(s) = \lim_{T\to\infty} \frac{1}{T}\, \mathbb{E}\left[\sum_{t=0}^{T-1} r_t \,\middle|\, s\right]$. If the MDP is unichain, then the average reward is state-independent, i.e., $\rho^\pi(s) \equiv \rho^\pi$ for any start state $s \in \mathcal{S}$ (Mahadevan, 1996a). The *relative value function* (or *bias*) from state s measures the difference between the actual rewards under π starting from s and the gain, $h^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \left(r_t - \rho^\pi\right) \,\middle|\, s\right]$.

A policy π^* is *optimal* in the average-reward case if *(i)* it has the highest possible gain, i.e., $\forall\, \pi,\ s : \rho^{\pi^*}(s) \geq \rho^\pi(s)$; and *(ii)* it has the highest bias with respect to the optimal gain, i.e., $\forall\, \pi,\ s : h^{\pi^*}(s) \geq h^\pi(s)$. An ϵ-optimal gain policy π^ϵ is a policy such that its gain at each state is at most ϵ worse than the optimal gain, i.e., $\rho^{\pi^*}(s) - \rho^{\pi^\epsilon}(s) \leq \epsilon\ \forall s$.

Advantages and Value Iteration Algorithm. A fundamental algorithm for solving MDPs, and the primary object of study in this paper, is the **Value Iteration** (VI) algorithm, which starts from an arbitrary value vector V_0 and then iteratively performs the update

$$V_{t+1}(s) \;=\; \max_{\mathrm{st}(z)=s} \left\{ r^z \;+\; \gamma \sum_{i=1}^{n} p_i^z\, V_t(s_i) \right\}. \tag{1}$$

Alternatively, VI can be expressed in terms of **advantages** (Mustafin et al., 2025), thereby simplifying the subsequent analysis. Thus, let us next introduce the concept of advantages. The advantage of a SAP $z = (s, a)$ with respect to a policy π is defined as the gain (or loss) caused by a one-time deviation from policy π by choosing action a at state s instead of following π starting from s. In the discounted-reward case, advantages are always well-defined and given by

$$\mathrm{adv}(z, \pi) \;=\; r^z \;+\; \gamma \sum_{i=1}^{n} p_i^z\, V^\pi(i) \;-\; V^\pi(s), \tag{2}$$

and this quantity exists for all SAPs and for all policies.

In the average-reward case, we define the advantage function as (Zhang & Ross, 2020; Adamczyk et al., 2025): $\mathrm{adv}(z, \pi) = Q^\pi(z) - h^\pi(s)$, where

$$Q^\pi(z) = \mathbb{E}\left[\sum_{t=0}^{\infty} \left(r_t - \rho^\pi\right) \mid s_0 = s, a_0 = a\right] = r^z - \rho^\pi + \sum_{i=1}^{n} p_i^z h^\pi(s_i).$$

Then,

$$\text{adv}(z, \pi) = r^z - \rho^\pi + \sum_{s'} p^z_{s'} h^\pi(s') - h^\pi(s).$$

Note that in both cases, we do not need the values to be produced by an actual policy to compute an advantage or that we can compute advantage with respect to any set of values, which allows to express VI in terms of advantages:

$$V_{t+1}(s) \;=\; V_t(s) \;+\; \max_{\text{st}(z)=s} \text{adv}(z, V_t), \tag{3}$$

and we will adopt this perspective here, as it aligns naturally with the geometric view.

Finally, we analyze the convergence of VI with respect to the vector span seminorm, defined as the difference between the maximum and minimum entries of a vector. For a value vector V, we denote it by $\text{sp}(V) = \max_i V(s_i) - \min_j V(s_j)$.

Geometric interpretation of MDPs. The discounted-reward MDP model presented above was reinterpreted in geometric terms by Mustafin et al. (2025), and our analysis relies on this interpretation. In this framework, key objects of an MDP—SAPs and policies—are viewed as points and hyperplanes in a linear space called the *SAP space*, which we make precise below. Consequently, the dynamics of VI can be interpreted as the movement of a hyperplane in this space.

In this paper, we extend this interpretation to the average-reward case. We show that, although the definition of values must be modified so that they still correctly define a hyperplane, the two settings are geometrically equivalent, which enables a unified analysis.

Before, we briefly introduce the geometric interpretation and define the quantities required for our analysis; we refer the reader to the original paper for a detailed exposition (Mustafin et al., 2025).

The SAP space is an $(n+1)$-dimensional space, where n is the number of states. One coordinate, referred to as the 0-th coordinate or the *height*, is special and corresponds to SAP rewards and policy values. To construct the **SAP vector** $z^+ = (c_0, \ldots, c_n)$ from a SAP z on state s ($\text{st}(z) = s$), we set its first entry (the 0-th coordinate) to the SAP reward: $c_0 = r^z$. The next n coordinates correspond to states and are defined as $c^z_i = \gamma p^z_i$ for $i \neq s$ and $c^z_s = \gamma p^z_s - 1$. Hence, the sum of the last n coordinates satisfies $\sum_{i=1}^n c^z_i = \gamma - 1$, with the s-th entry negative and all others positive.

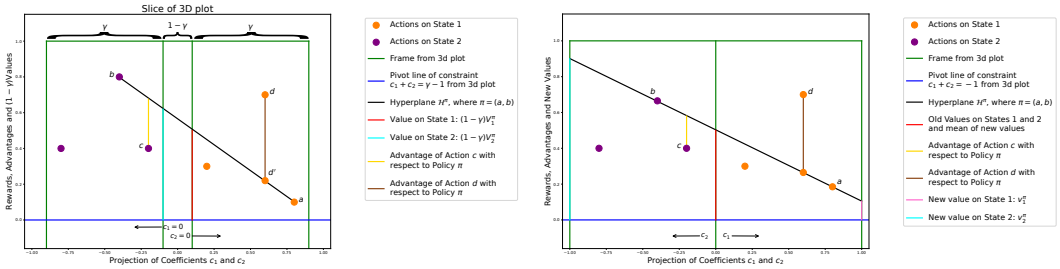
For a policy π, the **policy vector** $V^\pi_+ = (1, V^\pi(s_1), \ldots, V^\pi(s_n))^\top$ concatenates the state values with a 1 in the 0-th coordinate. Geometrically, π is represented by the hyperplane $\mathcal{H}^\pi$ consisting of all vectors orthogonal to V^π_+ (with respect to the standard Euclidean inner product), which is the span of the SAP vectors forming the policy. Such a hyperplane can be constructed for any value vector V_+, i.e., we do not require an explicit policy; for example, we can form a hyperplane for V_t produced by t iterations of VI. Hyperplanes arising from value vectors without associated policies are called **pseudo-policies**.

The key property of action and policy vectors is that for any SAP vector z^+ and policy vector V^π_+, the inner product $z^+ V^\pi_+$ equals the advantage of z under π and corresponds to the oriented vertical distance from z^+ to $\mathcal{H}^\pi$. Consequently, if SAP z is in policy π, the advantage of z with respect to π is 0 and vectors z^+ and V^π_+ are orthogonal. An action-space example for a two-state MDP is shown in Figure 2a.

One tool provided by the geometric interpretation that we use here is the equivalence transformation $\mathcal{L}^\delta_s$, which transforms an MDP $\mathcal{M}$ into a new MDP $\tilde{\mathcal{M}} = \mathcal{L}^\delta_s \mathcal{M}$. The transformation preserves transition probabilities, but modifies rewards of SAPs such that for any policy π the original values V^π in $\mathcal{M}$ and transformed values $\tilde{V}^\pi$ in $\tilde{\mathcal{M}}$ satisfy:

- $\tilde{V}^\pi(s) = V^\pi(s) + \delta$ for the transformed state s,
- $\tilde{V}^\pi(s') = V^\pi(s')$ for any state $s' \neq s$,
- the advantages $\text{adv}(z, \pi)$ of all SAPs remain unchanged.

This transformation allows us to obtain a **normalization** by applying n transformations, one per state: $\mathcal{L}^{-v^*(s_i)}_{s_i}$. The resulting MDP $\mathcal{M}^*$ has optimal values equal to 0 at every state, so the hyperplane $\mathcal{H}^{\pi^*}$ corresponding to the optimal policy becomes horizontal at level 0. Consequently, the



(a) 2-State MDP, old interpretation, discounted reward. (b) 2-State MDP, new interpretation, average reward.

Figure 2: Old (left) and new (right) visualizations of a two-state MDP with the same transition probabilities and rewards but different discount factors. The right panel also illustrates why the geometric interpretation developed for the discounted-reward case does not extend directly to the average-reward case: in the latter, the vertical value lines collapse into a single line, so all states have the same value, and this set of values does not define a unique hyperplane. At the same time, a hyperplane can still be constructed for an ergodic policy, and the geometric picture remains logically coherent in that case. To extend the geometric framework to the average-reward case, we propose measuring the values not on the inner but on the outer edges of the action zones. These new values v can be used in both the average- and discounted-reward cases.

rewards of SAPs participating in optimal policies are 0, while the rewards of all other SAPs are negative. The other properties of this normalization relevant to our analysis are discussed in the next section.

3 Unified Geometric Interpretation and New Policy Values

In the discounted-reward setting, we can rearrange the Bellman equation to obtain the linear system $(I - \gamma P^\pi)V = R^\pi$, which possesses a unique fixed point V^π as $I - \gamma P^\pi$ is invertible. However, this formula breaks down in the average-reward case $\gamma = 1$, where the matrix $I - P^\pi$ is singular, and the discounted value function is no longer well-defined. Instead, the analysis relies on the gain function ρ^π and the relative value function h^π. This can introduce two difficulties, *(i)* the discounted case and average-reward case require different value functions, and *(ii)* the relative value function is not unique, since $h^\pi + c\mathbb{1}$ is also a solution for any constant c.

This problem can also be seen in the geometric interpretation in Figure 2, where the value function $V^\pi(s)$ is represented by the inner vertical line on the state s. When $\gamma = 1$, the vertical lines corresponding to different states coincide, reflecting the degeneracy of the classical value representation in the average-reward case. Nevertheless, this observation suggests a potential avenue for a new interpretation. The advantage function remains well-defined for any value of γ, including $\gamma = 1$. We develop the geometric interpretation below. The additional algebraic and geometric intuition is provided in Appendix E.1.

New Action and Policy Vectors. As suggested by Figure 2b, we propose measuring values along the outer vertical lines, rather than along the inner vertical lines as previously, which leads to the following definitions of the action and policy vectors.

For SAP $z = (s, a)$, let the **new SAP vector** be $z^+ = \left(r^z, \frac{\gamma p_1^z - \gamma}{C}, \ldots, \frac{\gamma p_s^z - 1 - \gamma}{C}, \ldots, \frac{\gamma p_n^z - \gamma}{C}\right)$, where $C = n\gamma + (1 - \gamma)$ is an MDP constant. For convenience, denote by c_i^z the i-th coordinate of z^+ excluding the first element r^z, that is,

$$c_i^z = \begin{cases} \frac{\gamma p_i^z - \gamma}{C}, & \text{if } i \neq s, \\ \frac{\gamma p_s^z - 1 - \gamma}{C}, & \text{if } i = s. \end{cases}$$

Note that the sum of all entries of z^+ except the first one equals -1. The first entry is special; following Mustafin et al. (2025), we refer to it as the 0-th coordinate (or *height*).

Given a unichain policy π, define the state values $v^\pi(s)$ by

$$\frac{v^\pi(s)}{C} \;=\; r^{(s,\pi(s))} \;-\; \frac{\gamma\, v^\pi_\Sigma}{C} \;+\; \gamma \sum_{i=1}^{n} p_i^{(s,\pi(s))} \frac{v^\pi(s_i)}{C}, \quad \text{where } v^\pi_\Sigma = \sum_{i=1}^{n} v^\pi(s_i). \tag{4}$$

Then define the **new policy vector** v^π_+ as $v^\pi_+ \;=\; \big(1,\; v^\pi(s_1),\; \ldots,\; v^\pi(s_n)\big)$.

Below, we first discuss how this unified value representation leads to a consistent policy evaluation and a natural normalization of the MDP, such that we can verify the existence of the value and reflect the geometric inspiration in an algebraic way. We then relate these newly introduced quantities to the classical value definitions in both the discounted- and average-reward settings. Throughout this section, V^π denotes the classical value function, whereas v^π denotes the new value representation introduced here.

Policy Evaluation and MDP Normalization. In the average-reward case, the matrix $I - \gamma P$ is not invertible when $\gamma = 1$, so a unique value vector V^π satisfying $R^\pi = (I - \gamma P^\pi)V^\pi$ does not exist. However, rewriting Equation (4) in matrix form and rearranging terms yields $R^\pi = (I + \gamma E - \gamma P^\pi)\,\frac{v^\pi}{C}$, where $E = \mathbb{1}\mathbb{1}^\top$ is the all-ones matrix. Whenever the matrix $I + \gamma E - \gamma P^\pi$ is invertible, the solution admits the explicit representation $v^\pi = C(I + \gamma E - \gamma P^\pi)^{-1} R^\pi$. We can now give the algebraic definition of these new values:

Definition 3.1 (New value function)**.** For a given policy π, the proposed value vector $v^\pi \in \mathbb{R}^n$ is defined as the unique solution to the linear system

$$v^\pi = C(I + \gamma E - \gamma P^\pi)^{-1} R^\pi, \tag{5}$$

where $C = n\gamma + 1 - \gamma$ for $\gamma \in (0, 1]$.

This equation can be solved in the average-reward case (i.e., $\gamma = 1$) when the policy π induces a unichain Markov chain, as shown in the following lemma. The proof is provided in Appendix C.

Lemma 3.2. *Let π be a policy with transition kernel P^π on a finite state space, and let $\mathbb{1}$ denote the all-ones vector with $E := \mathbb{1}\mathbb{1}^\top$. The matrix $I + E - P^\pi$ is invertible if, and only if, the Markov chain induced by P^π is unichain. In particular, in this case, the values v^π defined via the linear system $(I + E - P^\pi)\, v^\pi = nR^\pi$ are uniquely determined.*

Since we can compute values for any unichain policy π, we can compute the advantage of any SAP with respect to such a policy, which is equivalent to constructing the policy hyperplane $\mathcal{H}^\pi$. This, in turn, implies that the transformation $\mathcal{L}^\Delta$ described by Mustafin et al. (2025) can be applied to the MDP. Therefore, if the optimal policy induces a unichain MDP (the case we analyze in this work), the MDP can be normalized, which significantly simplifies the analysis.

Note that **normalization is not an additional assumption**. It is always possible in the discounted-reward case and is possible in the average-reward case under the assumptions made in this work.

Let v^π_{orig} denote the values in the original MDP $\mathcal{M}_{\text{orig}}$ and let v^π denote the values in the normalized MDP $\mathcal{M}$. The normalization transformation (Mustafin et al., 2025) has two key properties with respect to VI dynamics (Theorem 3.5):

- Since it preserves advantages, the state-wise maximizing SAPs are preserved, and therefore the same greedy policy is selected at every VI iteration. The corresponding value iterates remain related by the same normalization transformation.
- Since it shifts all values in a state by the same quantity δ, differences between policy values are preserved, i.e., for any two policies π and π', $v^\pi - v^{\pi'} = v^\pi_{\text{orig}} - v^{\pi'}_{\text{orig}}$.

Together, these properties imply that if, after T iterations, VI produces an ϵ-optimal policy in the normalized MDP, then the same policy is also ϵ-optimal in the original MDP. Therefore, it is sufficient to analyze the normalized MDP instead of the original one.

Overall, normalization does not change the structure of the MDP; it only simplifies the analysis and makes the geometry more transparent. Algebraically, the analysis presented in this paper can be formulated in terms of the error vectors $e^\pi = v^\pi_{\text{orig}} - v^*_{\text{orig}}$, and the advantages with respect to the optimal policy, $\text{adv}(z, v^*)$, instead of the normalized values v^π and rewards r^z.

Advantage Representation under the New Values. The geometric interpretation motivates the design of the new value function so that the policy hyperplane remains unchanged and the advantage function is preserved. The following lemma states that under the proposed value representation, the inner product of the SAP vector and the policy vector recovers the classical advantage function.

Lemma 3.3. *For $\gamma \in (0, 1]$, the inner product of SAP vector z and policy vector π is equal to the advantage of SAP z with respect to π, i.e., $z^+ v_+^\pi = r^z + \sum_i c_i^z v^\pi(i) = adv(z, \pi)$.*

The result establishes equivalence between the old and new geometric interpretations in both the discounted and average-reward cases. Consequently, the geometric interpretation of policy hyperplanes remains valid and can be used directly in VI. The detailed derivations are provided in Appendix B, where we establish the connections between the new value function and the classical value function, the average reward, and the relative value function.

4 Main Result

Thus far, we have established the structural equivalence of the new value function in the discounted and average-reward cases via the geometric interpretation. Next, we investigate the algorithmic consequences. In particular, we analyze the convergence behavior of VI based on the advantage function associated with the new value function v^π and show geometric convergence in both the discounted and average-reward cases. For clarity of exposition, we consider the normalized MDP in the following and detailed proofs of the results in this section, and supporting lemmas are deferred to Appendix D. We work under the following assumption, whose feasibility is discussed in Section 5.

Assumption 4.1. There is a **unique** optimal policy π^*. The MDP implied by π^* is **unichain** and **aperiodic**.

Now, we present the **Value Iteration algorithm** on the new values v^π. Starting from the arbitrary values v_0, the algorithm consists of iteratively performing an update similar to Equation (1):

$$\frac{v_{t+1}}{C} = \frac{v_t}{C} + \max_z \text{adv}(z, v_t) = R^{\pi_t} - E\frac{v_{t\Sigma}}{C} + \gamma P^{\pi_t}\frac{v_t}{C},$$

where π_t is a policy greedy with respect to the values v_t, $\pi_t(i) = \arg\max_{z,\text{st}(z)=i} \text{adv}(z, v_t)$. The resulting algorithm, although it chooses a different anchor, shares similarities with the Relative Value Iteration algorithm (White, 1963), which inspired our choice of v as the notation for the new values.

Then, we show that the new update has a contraction property with respect to $\text{sp}(v_t)$, the span seminorm of the value vector.

Theorem 4.2. *If Assumption 4.1 holds, the span of the normalized value vector obtained after $T = n^2$ steps of a standard Value Iteration algorithm satisfies the following inequality:*

$$sp(v_T) \le \gamma^T \iota \, sp(v_0),$$

where $\gamma \in (0, 1],\ \iota \in (0, 1)$.

Here, ι is an MDP-dependent contraction factor ($\iota = 1 - \bar{\phi}|\mathcal{S}_R|$ in the proof), arising from the uniform positive contribution of the recurrent states of the optimal-policy-induced Markov chain over T iterations. The bound is primarily theoretical and is generally conservative; we discuss the sources of this conservativeness in Appendix E.2.

The contraction property holds for all $\gamma \in (0, 1]$, implying asymptotic convergence of value iteration in the span seminorm. For $\gamma \in (0, 1)$, corresponding to the discounted-reward setting, geometric convergence follows from the classical factor γ^T and is further strengthened by the additional factor ι. For $\gamma = 1$, corresponding to the average-reward setting, geometric contraction is induced entirely by ι.

Proof. We give the proof strategy here; the full proof is provided in Appendix D. The proof consists of six steps.

Step 1: Using the properties of the normalized MDP, we derive upper and lower bounds on each component of $\widetilde{v}_{t+1}$ using the optimal SAP and the current greedy SAP.

Step 2: We combine these bounds to represent the one-step update using a stochastic update matrix:

$$\widetilde{v}_{t+1} + \gamma E\widetilde{v}_t = \gamma P_t'\widetilde{v}_t, \qquad P_t' = D_t P^\star + (I - D_t)P_t.$$

Step 3: The strict advantage gap implied by uniqueness of the optimal policy gives a uniform positive lower bound on the entries of the diagonal matrix D_t, allowing us to decompose

$$P_t' = \widetilde{P}_t' + \delta' P^\star, \qquad \delta' > 0.$$

Step 4: We combine the updates over T iterations by introducing the T-step kernel

$$K_T = P_T' \cdots P_1',$$

and express each component of $\widetilde{v}_T$ as a weighted combination of the initial values, with coefficients $\lambda_i^s = [K_T]_{s,i}$.

Step 5: Using the unichain and aperiodicity properties of $P^\star$, we obtain a uniform positive lower bound on the coefficients λ_i^s associated with its recurrent states. Choosing a strictly smaller common lower bound $\bar{\phi} > 0$ yields a Doeblin minorization of the T-step kernel.

Step 6: Finally, we separate this common recurrent-state component from the remaining coefficients. The common component is state-independent and therefore cancels when taking the span, while the remaining coefficients have total weight $1 - \bar{\phi}|\mathcal{S}_R| < 1$. This yields the contraction factor

$$\iota = 1 - \bar{\phi}|\mathcal{S}_R| \in (0, 1).$$

□

The following corollaries extend the convergence guarantee obtained for the normalized MDP to the original MDP. Using the correspondence induced by the normalization transformation, they translate the span-convergence bound into ϵ-optimality guarantees and derive the resulting iteration complexity.

Corollary 4.3. *In the discounted reward case, the Value Iteration Algorithm outputs an ϵ-optimal policy after*

$$\mathcal{O}\left(\frac{\log(1/\epsilon) + \log(1/(1-\gamma))}{\log(1/\gamma) + \frac{\log(1/\iota)}{n^2}}\right)$$

iterations.

Corollary 4.4. *In the average-reward case, the Value Iteration Algorithm outputs an ϵ-optimal policy after*

$$\mathcal{O}\left(\frac{\log(1/\epsilon)}{\frac{\log(1/\iota)}{n^2}}\right)$$

iterations.

5 Discussion and Conclusion

Discussion. Comparison with Lee & Ryu (2025). Let us explicitly compare our results to the state-of-the-art analysis presented by Lee & Ryu (2025), in particular, their Theorem 3.

Theorem 5.1 (Theorem 3 from Lee & Ryu (2025))**.** *Let $t \geq 0$, $n \geq t+2$, and $V_0 \in \mathbb{R}^n$. Then there exists a unichain MDP with $|\mathcal{S}| = n$ and $|\mathcal{A}| = 1$ such that its modified Bellman equations has a solution (g^*, h^*) satisfying*

$$\left\| \sum_{i=0}^{t} \beta_i (TV_i - V_i) - g^* \right\|_\infty \geq \frac{1}{t+1} \|V_0 - h^*\|$$

for any iterates $\{V_i\}_{i=0}^t$ satisfying the span condition

$$V_{t+1} \in V_o + \mathit{span}\{TV_0 - V_0, \ldots, TV_k - V_k\}$$

and any choice of real numbers $\{\beta_i\}_{i=0}^k$ such that $\sum_{i=0}^k \beta_i = 1$.

Then, setting β_k to 1 gives the lower bound on the convergence of the VI algorithm.

The theorem is proved by providing an example for which the convergence of VI cannot exceed a sublinear rate. The example constructed in the proof is a unichain MDP that satisfies Assumption 4.1. At first glance, this seems to contradict our main result. Does this mean that one of the papers contains a technical error?

The answer is No. The apparent contradiction arises from two features of the analysis by Lee & Ryu (2025): the use of the ℓ^∞ norm of the Bellman error as the performance measure and the consideration of a small number of iterations relative to the number of states. For the former, Lee & Ryu (2025) also discuss convergence in terms of the span seminorm at the end of Section 2, where they observe that convergence in ℓ^∞ implies convergence in the span seminorm because the span seminorm is upper bounded by twice the ℓ^∞ norm. This bound shows that the span seminorm cannot converge *slower* than the ℓ^∞ norm, but it does not rule out that it can converge *significantly faster*. As we show in this paper, this is indeed the case under Assumption 4.1. Note that, although the infinity norm is commonly used in classical analyses, the span seminorm is sufficient here for deriving policy-quality guarantees, as shown in Corollaries 4.3 and 4.4, it suffices for evaluating the quality of the resulting policy.

The second feature concerns the number of algorithm iterations. Lee & Ryu (2025) establish the bound for $t \leq n-2$ iterations. This choice ensures that the sublinear rate follows from an information-based complexity argument. Intuitively, there exist two states in the graph such that the shortest path between them requires exactly $n-1$ steps. If the number of iterations is at most $n-2$, the first state cannot receive any information from the last state, and therefore the value error may remain arbitrarily large. As we show in this paper, n^2 iterations are sufficient for any two states to communicate, after which the geometric convergence rate becomes visible in finite time.

Taken together, these two features enable an analysis that is technically correct, but may lead to the misleading conclusion that VI exhibits only sublinear finite-time convergence in the unichain case. In contrast, our results show that when the optimal policy is unichain (which always holds in the unichain MDP setting), the convergence rate is in fact geometric.

Discussion on Assumption 4.1 and the multichain case. We now discuss the limitations of Assumption 4.1. At first glance, this assumption may appear restrictive; however, in the average-reward setting, it is often satisfied. Aperiodicity can be ensured by introducing self-loops at each state, for example, via a standard lazy transition (mixing with the identity) (Sec . 8.5.4, Puterman (1994)). Such a transformation does not change the policy gains in average-reward MDPs, and therefore does not affect the conclusions of the analysis. Then, the prerequisite required to guarantee uniqueness is a strict preference over states and actions, which can be interpreted as a tie-breaking rule (e.g., induced by an arbitrarily small perturbation of the rewards). Under this condition, the unichain optimal policy follows from a standard connectivity assumption commonly used in exploration analysis (Auer et al., 2008; Boone & Gaujal, 2025). This assumption states that for any two states s and s', there exists a policy π such that s' is reachable from s with positive probability in finite time, i.e.,

$$\forall s, s' \; \exists \pi : \; \mathbb{P}_\pi\big(\tau_{s'} < \infty \mid s_0 = s\big) > 0,$$

where $\tau_{s'} = \inf\{t \geq 0 : s_t = s'\}$ is the hitting time of s'. The connectivity assumption is consistent with the unichain setting and is satisfied in a broad class of MDPs where actions allow transitions between states. Thus, the assumption fails only when there exists at least one isolated class (i.e., a set of states that cannot be left under any policy) together with another class of connected states that achieves a higher gain under any policy. In this setting, the same line of analysis cannot be applied, since it requires sufficiently accurate value estimates to provide guarantees on the resulting policy. A general analysis of this case is left for future work.

Conclusion. We analyzed the convergence rates of VI in the discounted and average-reward settings. Our analysis revealed that, under a unique, unichain optimal policy, VI converges geometrically with a rate strictly faster than γ in the span seminorm. This result significantly improves upon earlier analysis, which suggested that sublinear convergence is optimal in the average-reward setting. The analysis was based on a geometric interpretation of MDPs, which enabled us to analyze both settings jointly, in contrast to most prior work that analyzes them separately.

Acknowledgements

This work was supported by the Research Council of Finland Flagship programme: Finnish Center for Artificial Intelligence FCAI. We acknowledge the financial support of the Finnish Ministry of Education and Culture through the Intelligent Work Machines Doctoral Education Pilot Program (IWM VN/3137/2024-OKM-4).

References

Jacob Adamczyk, Volodymyr Makarenko, Stas Tiomkin, and Rahul V Kulkarni. Average-reward soft actor-critic. In *Reinforcement Learning Conference*, 2025.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

R. Bellman. *Dynamic Programming.* Dover Publications, 1957.

David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33:719–726, 1962.

Victor Boone and Bruno Gaujal. Logarithmic regret of exploration in average reward markov decision processes. *arXiv preprint arXiv:2502.06480*, 2025.

Eric V Denardo and Bennett L Fox. Multichain markov renewal programs. *SIAM Journal on Applied Mathematics*, 16(3):468–487, 1968.

Awi Federgruen and Paul J. Schweitzer. The functional equations of undiscounted markov renewal programming. *Mathematics of Operations Research*, 3(3):308–321, 1978.

Eugene A Feinberg and Gaojin He. Complexity bounds for approximately solving discounted mdps by value iterations. *Operations Research Letters*, 48(5):543–548, 2020.

Eugene A Feinberg and Jefferson Huang. The value iteration algorithm is not strongly polynomial for discounted dynamic programming. *Operations Research Letters*, 42(2):130–131, 2014.

Julien Grand-Clément and Marek Petrik. Reducing blackwell and average optimality to discounted mdps via the blackwell discount factor. *Advances in Neural Information Processing Systems*, 36: 52628–52647, 2023.

Ronald A Howard. *Dynamic programming and markov processes.* John Wiley, 1960.

Gerhard Hübner and Karl Hinderer. On exact and approximate solutions of unstructured finite-stage dynamic programs. In H. C. Tijms and J. Wessels (eds.), *Markov Decision Theory*, volume 93 of *Mathematical Centre Tracts*, pp. 57–76. Mathematical Centre, Amsterdam, 1977.

Jongmin Lee and Ernest K. Ryu. Accelerating value iteration with anchoring. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.16569.

Jongmin Lee and Ernest K Ryu. Optimal non-asymptotic rates of value iteration for average-reward mdps. In *13th International Conference on Learning Representations, ICLR 2025*, pp. 93671–93713. International Conference on Learning Representations, ICLR, 2025.

Sridhar Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22(1):159–195, 1996a.

Sridhar Mahadevan. Sensitive discount optimality: Unifying discounted and average reward reinforcement learning. In *ICML*, pp. 328–336, 1996b.

Thomas E. Morton and William E. Wecker. Discounting, ergodicity and convergence for markov decision processes. *Management Science*, 23(8):890–900, 1977.

Arsenii Mustafin, Aleksei Pakharev, Alex Olshevsky, and Ioannis Ch Paschalidis. Mdp geometry, normalization and reward balancing solvers. *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics, PMLR 258:2476-2484*, 2025.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, NY, 1994. ISBN 978-0471619772. DOI: 10.1002/9780470316887.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley, Hoboken, NJ, 2014. Reprint of the 1994 edition.

Paul J Schweitzer. On the existence of relative values for undiscounted markovian decision processes with a scalar gain rate. *Journal of mathematical analysis and applications*, 104(1):67–78, 1984.

Paul J. Schweitzer and Awi Federgruen. The asymptotic behavior of undiscounted value iteration in markov decision problems. *Mathematics of Operations Research*, 2(4):360–381, 1977.

Paul J Schweitzer and Awi Federgruen. The functional equations of undiscounted markov renewal programming. *Mathematics of Operations Research*, 3(4):308–321, 1978.

Paul J. Schweitzer and Awi Federgruen. Geometric convergence of value-iteration in multichain markov decision problems. *Advances in Applied Probability*, 11(1):188–217, 1979.

E. Seneta. *Non-negative Matrices and Markov Chains*. Springer Series in Statistics. Springer, New York, 2nd edition, 2006.

Csaba Szepesvári. Value iteration and our first lower bound. RL Theory: Planning in MDPs, Lecture 3, 2021. URL `https://rltheory.github.io/lecture-notes/planning-in-mdps/lec3/`. Course notes for CMPUT 653: Theoretical Foundations of Reinforcement Learning.

John N Tsitsiklis and Benjamin Van Roy. On average versus discounted reward temporal-difference learning. *Machine Learning*, 49(2):179–191, 2002.

J. van der Wal. *Stochastic Dynamic Programming: Successive Approximations and Nearly Optimal Strategies for Markov Decision Processes and Markov Games*. Number 139 in Mathematical Centre Tracts. Mathematical Centre, Amsterdam, 1981.

Eugenio Della Vecchia, Silvia Di Marco, and Alain Jean-Marie. Illustrated review of convergence conditions of the value iteration algorithm and the rolling horizon procedure for average-cost mdps. *Annals of Operations Research*, 199(1):193–214, 2012. DOI: 10.1007/s10479-012-1070-0.

D. J. White. Dynamic programming, markov chains, and the method of successive approximations. *Journal of Mathematical Analysis and Applications*, 6(3):373–376, 1963.

AA Yushkevich. On a class of strategies in general markov decision models. *Theory of Probability & Its Applications*, 18(4):777–779, 1974.

Yiming Zhang and Keith W Ross. Average reward reinforcement learning with monotonic policy improvement. 2020.

A Details of the MDP example.

MDP construction (heavenpurgatoryhell). We consider a finite MDP $M = (\mathcal{S}, \mathcal{A}, P, r)$. The state space

$$\mathcal{S} = \{H_1, H_2, H_3,\ P_1, P_2, P_3,\ L_1, L_2, L_3\}$$

is partitioned into three groups: "heaven" $\mathcal{S}_H = \{H_1, H_2, H_3\}$, "purgatory" $\mathcal{S}_P = \{P_1, P_2, P_3\}$, and "hell" $\mathcal{S}_L = \{L_1, L_2, L_3\}$. We analyze the dynamics under a fixed (optimal) stationary policy $\pi^\star$ that selects a single available action in each state (the construction below specifies the induced transition kernel $P^{\pi^\star}$ and rewards).

Rewards. The perstep reward function is

$$r(s) = \begin{cases} 1, & s \in \mathcal{S}_H, \\ 0, & s \in \mathcal{S}_P, \\ -1, & s \in \mathcal{S}_L. \end{cases}$$

Transitions under $\pi^\star$. The induced rowstochastic matrix $P \equiv P^{\pi^\star}$ is specified by the following nonzero probabilities.

Heaven cycle with rare leakage to purgatory:

$$\begin{aligned} &\Pr(H_2 \mid H_1) = 0.75, \qquad \Pr(P_1 \mid H_1) = 0.25, \\ &\Pr(H_3 \mid H_2) = 0.75, \qquad \Pr(P_1 \mid H_2) = 0.25, \\ &\Pr(H_1 \mid H_3) = 0.75, \qquad \Pr(P_1 \mid H_3) = 0.25. \end{aligned}$$

Purgatory chain that typically returns to heaven, with leakage to hell:

$$\begin{aligned} &\Pr(P_1 \mid P_1) = 0.20, \quad \Pr(P_2 \mid P_1) = 0.60, \quad \Pr(L_1 \mid P_1) = 0.20, \\ &\Pr(P_2 \mid P_2) = 0.20, \quad \Pr(P_3 \mid P_2) = 0.60, \quad \Pr(L_1 \mid P_2) = 0.20, \\ &\Pr(P_3 \mid P_3) = 0.20, \quad \Pr(H_1 \mid P_3) = 0.60, \quad \Pr(L_1 \mid P_3) = 0.20. \end{aligned}$$

Hell cycle that occasionally escapes to purgatory:

$$\begin{aligned} &\Pr(P_2 \mid L_1) = 0.50, \qquad \Pr(L_2 \mid L_1) = 0.50, \\ &\Pr(P_2 \mid L_2) = 0.50, \qquad \Pr(L_3 \mid L_2) = 0.50, \\ &\Pr(P_2 \mid L_3) = 0.50, \qquad \Pr(L_1 \mid L_3) = 0.50. \end{aligned}$$

Summary. Under $\pi^\star$, the heaven states form a 3-cycle with reward 1 and a 0.25 chance per step to fall into purgatory; purgatory forms a directed chain that either advances, selfloops with probability 0.20, leaks to hell with probability 0.20, or returns to heaven from P_3 with probability 0.60; hell is a 3cycle with reward -1 and a 0.50 chance per step to move to P_2. This fully specifies $(\mathcal{S}, P, r)$ for the singleaction (policyfixed) analysis of this example.

B Relationship between the New Value and Classical Value

In this appendix, we provide the technical details underlying the connections between the proposed value function and the classical formulations in both the discounted and the average-reward settings. We first present the main relationships and then provide the proofs of the lemmas.

Discounted Reward Case From a geometric perspective, the difference between the new values and the classical values is the different locations of the vertical value lines on the states, as shown in Figure 2a and Figure 2b. To illustrate the algebraic interpretation of the new value function, we show the relationships between the old and new values in the discounted reward case.

When $\gamma < 1$ for the discounted reward case, we first construct the connection with the old value function.

Lemma B.1. *If $\gamma < 1$, the for a given π, there exists*

$$V^\pi(s) = \frac{v^\pi(s)}{C} + \frac{\gamma v^\pi_\Sigma}{C(1-\gamma)}. \tag{6}$$

Consequently, the sums of the value function exists as

$$v^\pi_\Sigma = (1-\gamma)V^\pi_\Sigma, \tag{7}$$

where $V^\pi_\Sigma = \sum_s V^\pi(s)$, and for the mean value $\bar{V}^\pi := \frac{V^\pi_\Sigma}{n}$ and $\bar{v}^\pi := \frac{v^\pi_\Sigma}{n}$, there exists

$$V^\pi(s) - \bar{V}^\pi = \frac{1}{C}(v^\pi(s) - \bar{v}^\pi).$$

This Lemma establishes the connection between the new value and the classical value in both component-wise and sum-wise forms. From the sum perspective, the factor $1-\gamma$ is introduced into the equivalence, which coincides with the geometric interpretation that, in the classical case, there exists a scaling weight $1-\gamma$ that reflects the value function on the vertical axis, whereas in the new case this scaling weight is removed.

With this, we can finally establish the equivalence between the old and new geometric interpretations in the discounted case.

Lemma B.2. *In the discounted case, the inner product of new SAP vector z^+ and new policy vector v^π_+ is equal to the advantage of SAP z with respect to values v^π:*

$$z^+ v^\pi_+ = r^z + \sum_i c^z_i v^\pi(i) = adv(z, \pi).$$

Thus, in the discounted case, the classic value iteration algorithm can still be implemented as the new value function does not change the advantage function. Furthermore, in our newly defined value function, γ can be 1, which motivates investigating equivalence in the average-reward case.

Average Reward Case. The equivalence of the new value function and the old value function in the discounted case has been established. We recall the equivalence between the discounted value function and the relative value function from the literature in the following lemma.

Lemma B.3 (Corollary 8.2.4 Puterman (1994))**.** *Given any policy π and state s,*

$$V^\pi(s) \;=\; \frac{\rho^\pi(s)}{1-\gamma} \;+\; h^\pi(s) \;+\; f^\pi(s,\gamma),$$

where V^π and h^π are the old discounted value function and the old relative value function, respectively; $\rho^\pi(s)$ is the average reward or gain, which is independent of s under standard unichain assumptions; and f^π is a remainder term such that $\lim_{\gamma\to 1} f^\pi(s,\gamma) = 0$.

As a result of $v^\pi_\Sigma = (1-\gamma)V^\pi_\Sigma$ shown in Lemma B.1 for the discounted case, we can get the equivalence between the traditional average reward and the new defined value function when considering the average-reward MDP.

Lemma B.4. *If $\gamma = 1$, then for a given π,*

$$\rho^\pi = \frac{v^\pi_\Sigma}{n}.$$

Lemma B.4 establishes the connection between the new value function and the average reward. In this way, the value function is still valid when applying $\gamma = 1$, and the average reward can be explicitly expressed by the value.

Let $\gamma = 1$ in our new value function in Equation (4). Since $C = n$, Lemma B.4, yields

$$\frac{v^\pi(s)}{C} \;+\; \rho^\pi \;=\; r^{\pi(s)} \;+\; \sum_{i=1}^{n} p_i^{\pi(s)} \frac{v^\pi(i)}{C}.$$

This new value function with MDP constant C can satisfy the Bellman equation in the average reward MDP:

$$T^\pi \frac{v^\pi(s)}{C} \;=\; r^{\pi(s)} \;+\; \sum_{i=1}^{n} p_i^{\pi(s)} \frac{v^\pi(i)}{C} = \frac{v^\pi(s)}{C} \;+\; \rho^\pi.$$

Our new value function $\frac{v^\pi(s)}{C}$ is unique to an additive constant, hence $\frac{v^\pi(s)}{C} \in \{h^\pi + c\mathbb{1} \mid c \in \mathbb{R}\}$. We then build an exact relationship between the new value function and the old relative value function.

Lemma B.5. *For a given policy π, there exists*

$$h^\pi(s) = \frac{v^\pi(s)}{n} - \frac{\rho^\pi}{n},$$

with $\sum_s h^\pi(s) = 0$.

In the average-setting, the bias function is defined only up to an additive constant, since if h^π is a solution of the Bellman equation, then $h^\pi + c\mathbb{1}$ is also a solution for any constant c. In contrast, the proposed value v^π is defined via the linear system and uniquely determined, as shown in Lemma 3.2. The relation in Lemma B.5 therefore identifies the unique representative of the affine solution set $\{h^\pi + c\mathbb{1} \mid c \in \mathbb{R}\}$ that satisfies the condition $\sum_s h^\pi(s) = 0$.

We now turn to the action-level characterization. In particular, we show that the inner product between the SAP vector and the policy vector recovers the advantage function under policy π in the average-reward setting.

Lemma B.6. *In the average reward case, the inner product of the SAP vector z^+ and the policy vector v_+^π is equal to the advantage of SAP z with respect to policy π:*

$$z^+ v_+^\pi = r^z + \sum_i c_i^z v^\pi(i) = \mathit{adv}(z, \pi).$$

In this section, we proposed that the newly defined value function remains well-defined in both the discounted and average-reward setting. Under the unichain assumption, the vector v^π/n satisfies the average-setting Bellman equation and therefore corresponds to the canonical bias function. Moreover, the geometric inner product between SAP vectors and policy vectors coincides exactly with the classical advantage function. This established the consistency of the proposed value function across the discounted and average reward cases.

In the following, we provide the proofs of the lemmas in this section.

Proof of Lemma B.1. If $\gamma < 1$, we can subtract the quantity $\frac{\gamma v_\Sigma^\pi}{C(\gamma-1)}$ from both sides of the value equation Equation 4, which leads to:

$$\begin{aligned}
\frac{v^\pi(s)}{C} - \frac{\gamma v_\Sigma^\pi}{C(\gamma-1)} &= r^{\pi(s)} - \frac{\gamma v_\Sigma^\pi}{C} - \frac{\gamma v_\Sigma^\pi}{C(\gamma-1)} + \gamma \sum_i p_i^{\pi(i)} \frac{v^\pi(i)}{C} \\
&= r^{\pi(s)} - \frac{\gamma v_\Sigma^\pi((\gamma-1)+1)}{C(\gamma-1)} + \gamma \sum_i p_i^{\pi(i)} \frac{v^\pi(i)}{C} \\
&= r^{\pi(s)} + \gamma \sum_i p_i^{\pi(i)} \left(\frac{v^\pi(i)}{C} - \frac{\gamma v_\Sigma^\pi}{C(\gamma-1)} \right).
\end{aligned}$$

Therefore, the values $\frac{v^\pi(s)}{C} - \frac{\gamma v_\Sigma^\pi}{C(\gamma-1)}$ satisfy the classic Bellman equation and, due to uniqueness,

$$V^\pi(s) = \frac{v^\pi(s)}{C} - \frac{\gamma v_\Sigma^\pi}{C(\gamma-1)} = \frac{(1-\gamma)v^\pi(s)}{C(1-\gamma)} + \frac{\gamma v_\Sigma^\pi}{C(1-\gamma)},$$

which indicates the relationship between the old value function V^π and the new defined value v^π in the discounted case. Then, we can sum this expression over all states to get the sum relationship:

$$\sum_s V^\pi(s) = \sum_s \left[\frac{(1-\gamma)v^\pi(s)}{C(1-\gamma)} + \frac{\gamma v_\Sigma^\pi}{C(1-\gamma)} \right] = \frac{(1-\gamma)v_\Sigma^\pi}{C(1-\gamma)} + \frac{\gamma n v_\Sigma^\pi}{C(1-\gamma)} = \frac{v_\Sigma^\pi}{1-\gamma}.$$

Defining the mean values as $\bar{V}^\pi = (\sum_i V^\pi(s))/n$ and similarly for $\bar{v}^\pi$, we obtain

$$\begin{aligned} V^\pi(s) - \bar{V}^\pi = V^\pi(s) - \frac{\sum_i V^\pi(i)}{n} &= \frac{(1-\gamma)v^\pi(s)}{C(1-\gamma)} + \frac{\gamma v_\Sigma^\pi}{C(1-\gamma)} - \frac{v_\Sigma^\pi}{(1-\gamma)n} \\ &= \frac{v^\pi(s)}{C} + \frac{\gamma n v_\Sigma^\pi}{Cn(1-\gamma)} - \frac{v_\Sigma^\pi(n\gamma + (1-\gamma))}{Cn(1-\gamma)} = \\ &= \frac{1}{C}(v^\pi(s) - \bar{v}^\pi). \end{aligned}$$

□

Proof of Lemma B.2. We work with the expression $\sum_i -c_i^z v^\pi(i)$, which is the height at which the hyperplane $\mathcal{H}^\pi$ crosses the vertical line with coordinates the same as z,

$$\begin{aligned} \sum_i -c_i^z v^\pi(i) &= \sum_i -c_i^z v^\pi(i) + \bar{v}^\pi - \bar{v}^\pi = \bar{v}^\pi + \sum_i -c_i^z (v^\pi(i) - \bar{v}^\pi) \\ &= (1-\gamma)\bar{V}^\pi + C\sum_i -c_i^z (V^\pi(i) - \bar{V}^\pi) \\ &= (1-\gamma)\bar{V}^\pi + (V^\pi(s) - \bar{V}^\pi) + \sum_i (\gamma - \gamma p_i^z)(V^\pi(i) - \bar{V}^\pi) \\ &= -\sum_i \tilde{c}_i^z V^\pi(i). \end{aligned}$$

□

Proof of Lemma B.4. From Lemma B.3 we have

$$v_\Sigma^\pi = (1-\gamma)\sum_s V^\pi(s) = \sum_s \rho^\pi(s) + (1-\gamma)\sum_s h^\pi(s) + (1-\gamma)\sum_s f^\pi(s,\gamma).$$

If $\gamma = 1$, the last two terms converge to 0. Then $v_\Sigma^\pi = \sum_s \rho^\pi$ and $C = |S| = n$. Therefore,

$$\rho^\pi = \frac{v_\Sigma^\pi}{n} = \frac{v_\Sigma^\pi}{C}.$$

□

Proof of Lemma B.5. For the relative function $h^\pi(s)$, by the equivalence from Lemma B.3 and the equivalence for value function in discounted MDP cases $V^\pi(s) = \frac{v^\pi(s)}{C} - \frac{\gamma v_\Sigma^\pi}{C(\gamma-1)}$, there is

$$\begin{aligned} h^\pi(s) &= V^\pi(s) - \frac{\rho^\pi(s)}{1-\gamma} - f^\pi(s,\gamma) \\ &= \frac{v^\pi(s)}{C} - \frac{\gamma v_\Sigma^\pi}{C(\gamma-1)} - \frac{\rho^\pi(s)}{1-\gamma} - f^\pi(s,\gamma) \\ &= \frac{v^\pi(s)}{C} + \frac{\frac{\gamma v_\Sigma^\pi}{C} - \rho^\pi(s)}{1-\gamma} - f^\pi(s,\gamma) \end{aligned} \tag{8}$$

When taking the limit for $\gamma \to 1$, we need to think about the convergence rate for the term $\frac{\gamma v_\Sigma^\pi}{C}$ as $C = n\gamma + 1 - \gamma$ is still a function of γ. Define $\phi(\gamma) = \frac{\gamma v_\Sigma^\pi}{n\gamma+1-\gamma}$. According to Taylor's expansion around 1, there is

$$\phi(\gamma) = \phi(1) + \phi'(1)(\gamma - 1) + \mathcal{O}(\gamma - 1),$$

where $\phi(1) = \frac{v_\Sigma^\pi}{n} = \rho^\pi$ and $\phi'(1) = \frac{(n\gamma+1-\gamma)v_\Sigma^\pi - \gamma(n-1)v_\Sigma^\pi}{(n\gamma+1-\gamma)^2} = \frac{v_\Sigma^\pi}{n^2}$. Plugging this expansion into (8) for $\gamma \to 1$, the equation reduces to

$$h^\pi(s) = \frac{v^\pi(s)}{n} + \lim_{t\to 1} \frac{\phi'(1)(\gamma-1) + \mathcal{O}(\gamma-1)}{1-\gamma}.$$

Since $\lim_{\gamma\to 1}\frac{\mathcal{O}(\gamma-1)}{1-\gamma}=0$, we obtain $h^\pi(s)=\frac{v^\pi(s)}{n}-\phi'(1)=\frac{v^\pi(s)}{n}-\frac{\rho^\pi}{n}$. $\phi'(1)$ is a constant and $\sum_s h^\pi(s)=0$.

□

Proof of Lemma B.6. For $\gamma=1$ we have $C=n$ and $\frac{v^\pi(s)}{n}=h^\pi(s)+\frac{\rho^\pi}{n}$. Let $c:=\frac{\rho^\pi}{C}$, then,

$$\begin{aligned}
z^+v_+^\pi &= r^z+\sum_i c_i^z v^\pi(i)\\
&= r^z-\frac{v_\Sigma^\pi}{C}+\sum_{i=1}^n p_i^z\frac{v^\pi(i)}{C}-\frac{v^\pi(s)}{C}\\
&= r^z-\rho^\pi+\sum_{i=1}^n p_i^z\left(h^\pi(i)+c\right)-\left(h^\pi(s)+c\right)\\
&= r^z-\rho^\pi+\sum_{i=1}^n p_i^z h^\pi(i)-h^\pi(s)\\
&= \mathrm{adv}(z,\pi).
\end{aligned}$$

□

C Proof for Section 3

Proof of Lemma 3.2. ($\Rightarrow$) Assume the Markov chain is unichain. Then, $\ker(I-P^\pi)=\mathrm{span}\{\mathbb{1}\}$ for the (right) kernel. For any vector x, write $x=\alpha\mathbb{1}+y$ with $\mathbb{1}^\top y=0$. Using $P^\pi\mathbb{1}=\mathbb{1}$ and $Ey=(\mathbb{1}^\top y)\mathbb{1}=0$,

$$(I+E-P^\pi)(\alpha\mathbb{1}+y)=\underbrace{(I-P^\pi)\alpha\mathbb{1}}_{=0}+\underbrace{E\alpha\mathbb{1}}_{=\alpha n\mathbb{1}}+(I-P^\pi)y+\underbrace{Ey}_{=0}=\alpha n\mathbb{1}+(I-P^\pi)y.$$

If $(I+E-P^\pi)(\alpha\mathbb{1}+y)=0$, left-multiplying by μ^T gives $\alpha n=0$, hence $\alpha=0$. Thus, $(I-P^\pi)y=0$ with $\mathbb{1}^\top y=0$. By $\ker(I-P^\pi)=\mathrm{span}\{\mathbb{1}\}$, we get $y=0$. Hence, $x=0$ is the only solution, so $I+E-P^\pi$ is invertible.

($\Rightarrow$) Assume $I+E-P^\pi$ is invertible. Suppose, for contradiction, that the chain is not unichain. Then there are at least two closed irreducible classes, and hence $\dim\ker(I-P^\pi)\geq 2$. Therefore, there exists a nonzero $y\in\ker(I-P^\pi)$ with $\mathbb{1}^\top y=0$ (choose a nontrivial linear combination of basis vectors of the kernel with zero sum). For this y, we have $Ey=0$ and

$$(I+E-P^\pi)y\;=\;(I-P^\pi)y\;+\;Ey\;=\;0,$$

which contradicts invertibility. Hence, the chain must be unichain.

The final statement follows since invertibility of $I+E-P^\pi$ guarantees a unique solution v^π to the corresponding linear system. □

D Proofs for Section 4

In this section, we provide the proofs of Theorem 4.2 and Corollaries 4.3 and 4.4. We precede the proof of the theorem with the following lemma.

Lemma D.1. *For any row-stochastic matrices $P,P_1,\cdots,P_T\in\mathbb{R}^{n\times n}$ and the all-ones matrix $E\in\mathbb{R}^{n\times n}$, there is*

$$\prod_{t=1}^T(P_t-E)=\prod_{t=1}^T P_t+\tilde{E},$$

where $\tilde{E}$ is a matrix with identical rows and $\tilde{E}v$ is a vector with identical entries for any vector v.

Proof of Lemma D.1. We first consider any row-stochastic matrix P, any matrix H whose rows are all identical and equal to a row vector h, and any vector v. Then:

- $PE = E$.
- EP is a matrix with identical rows. Moreover, writing $H = [\,h, \ldots, h\,]^\top$ with h a row vector, we have $Hv = [\,hv, \ldots, hv\,]^\top$, i.e., a vector whose entries are all the same scalar hv.
- For any $m \in \mathbb{N}$, $E^m = n^m E$. Hence, $E^m v$ is a vector with identical entries.

Thus, for any $i_1, i_2, \cdots, i_k \in \{1, 2, \cdots, n\}$, we have

- $P_{i_1} P_{i_2} \cdots P_{i_k} E = E$ as $PE = E$.
- The matrix $EP_{i_1} P_{i_2} \cdots P_{i_k}$ has identical rows. We note $EP_{i_1} = [h, \cdots, h]^\top$ since it has identical rows h. Then, for $P_{i_2} = [v_1, \cdots, v_n]$ with column vectors $v_1, \cdots, v_n$, $EP_{i_1} P_{i_2} = [h, \cdots, h]^\top [v_1, \cdots, v_n]$ shows that $EP_{i_1} P_{i_2}$ has identical rows $[hv_1, hv_2, \cdots, hv_n]$. By this induction, we obtain the result.
- $P_{i_1} P_{i_2} \cdots P_{i_j} E P_{i_{j+1}} P_{i_{j+2}} \cdots P_{i_k}$ has identical rows.

The product of matrices $\prod_{t=1}^{T} (P_t - E)$ can be expanded as

$$\prod_{t=1}^{T} (P_t - E) = \sum_{S \subseteq \{1,\ldots,T\}} (-1)^{|S|} \left(\prod_{t \in S} E \prod_{t \notin S} P_t \right),$$

where each subset $S \subseteq \{1, \ldots, T\}$ indicates the indices where E is selected instead of P_t. The first term corresponds to $\prod_{t=1}^{T} P_t$, and the following terms

$$\sum_{k=1}^{T} (-1)^k \sum_{1 \le i_1 < \cdots < i_k \le T} P_T \cdots P_{i_k+1} E P_{i_k-1} \cdots P_{i_1+1} E P_{i_1-1} \cdots P_1$$

contain one or more E factors. The product of E with any stochastic matrices from the left can be simplified to E. Consequently, each term in the expansion that contains k matrices E can be reduced to E^k multiplied by the remaining right-hand stochastic matrices, which all have identical rows as shown above. Moreover, the sum of matrices with identical rows also has identical rows, denoted as $\tilde{E}$.

□

With Lemma D.1, we are ready to prove the main theorem:

If Assumption 4.1 holds, the span of the normalized value vector obtained after $T = n^2$ steps of a standard Value Iteration algorithm satisfies the following inequality:

$$\mathrm{sp}(v_T) \le \gamma^T \iota \, \mathrm{sp}(v_0),$$

where $\gamma \in (0, 1],\ \iota \in (0, 1)$.

Proof of Theorem 4.2. **Step 1:** First, note that we analyze a normalized MDP. Therefore, all actions lie on or below the horizontal zero-level hyperplane or, in algebraic terms, SAPs participating in the unique optimal policy π^* have 0 rewards, and SAPs that do not participate have negative rewards.

For clarity, we define $\tilde{v}_t = \frac{v_t}{C}$. With a slight abuse of notation, we write

$$\mathrm{adv}(z, \tilde{v}_t) := r^z + \gamma \sum_i (p_i^z - 1) \tilde{v}_t(i) - \tilde{v}_t(s),$$

which equals $\mathrm{adv}(z, v_t)$ under the original parameterization. Denote the SAP chosen by the optimal policy in the state s as z^* and the SAP greedy with respect to the values $\tilde{v}_t$ as z ($z^* \in \pi^*$, $z := \arg\max_{z'} \mathrm{adv}(z', \tilde{v}_t)$ with $st(z^*) = st(z) = s$). Then, we have

$$\tilde{v}_{t+1}(s) = \tilde{v}_t(s) + \mathrm{adv}(z, \tilde{v}_t) = r^z + \gamma \sum_i (p_i^z - 1) \tilde{v}_t(i) \le \gamma \sum_i (p_i^z - 1) \tilde{v}_t(i),$$

since all rewards in a normalized MDP are non-positive. Moreover,

$$\tilde{v}_{t+1}(s) = \tilde{v}_t(s) + \text{adv}(z, \tilde{v}_t) \geq \tilde{v}_t(s) + \text{adv}(z^*, \tilde{v}_t) = \gamma \sum_i (p_i^{z^*} - 1)\tilde{v}_t(i),$$

since the rewards of all optimal SAPs in a normalized MDP are 0.

Step 2: Combining these bounds, we obtain

$$\begin{aligned}
&\gamma \sum_i (p_i^{z^*} - 1)\tilde{v}_t(i) \;\leq\; \tilde{v}_{t+1}(s) \;\leq\; \gamma \sum_i (p_i^{z} - 1)\tilde{v}_t(i), \\
&\Rightarrow \gamma(P^* - E)\tilde{v}_t \;\leq\; \tilde{v}_{t+1} \;\leq\; \gamma(P_t - E)\tilde{v}_t, \\
&\Rightarrow \gamma P^* \tilde{v}_t \;\leq\; \tilde{v}_{t+1} + \gamma E \tilde{v}_t \;\leq\; \gamma P_t \tilde{v}_t, \\
&\Rightarrow \tilde{v}_{t+1} + \gamma E \tilde{v}_t = \gamma P_t' \tilde{v}_t = \gamma \left[D_t P^* + (I - D_t) P_t \right] \tilde{v}_t,
\end{aligned}$$

where E is the all-ones matrix and $P_t' := D_t P^* + (I - D_t)P_t$ represents a coordinate-wise convex combination of the upper and lower bounds. The diagonal matrix $D_t \in \mathbb{R}^{n \times n}$ has the weights of this combinations, $D_t(s, s) \in [0, 1]$. When the optimal action is chosen, the two bounds coincide, and we set $D_t(s, s) = 1$. Besides, $E\tilde{v}_t$ is a vector with same entries, therefore for $\gamma \leq 1$,

$$\text{sp}(\tilde{v}_{t+1}) = \gamma \text{sp}(P_t' \tilde{v}_t) \leq \gamma \text{sp}(\tilde{v}_t) \leq \text{sp}(\tilde{v}_t), \tag{9}$$

where the first inequality holds because, for any state s there exists $\min_i \tilde{v}_t(i) \;\leq\; P_t' \tilde{v}_t(s) \;\leq\; \max_i \tilde{v}_t(i)$.

Step 3: Define the maximum advantage of non-optimal actions with respect to the optimal policy as $-\delta$:

$$-\delta = \max_{z' \notin \pi^*} \text{adv}(z', \pi^*), \quad \delta > 0.$$

Hence,

$$\text{adv}(z', \pi^*) \leq -\delta < 0, \quad \forall z' \notin \pi^*.$$

Since $r^z = \text{adv}(z, \pi^*) \leq -\delta < 0$, we obtain a stricter upper bound for $\tilde{v}_{t+1}(s)$ for a state s in which a non-optimal action is chosen:

$$\tilde{v}_{t+1}(s) = r^z + \gamma \sum_i (p_i^z - 1)\tilde{v}_t(i) \leq -\delta + \gamma \sum_i (p_i^z - 1)\tilde{v}_t(i) < \gamma \sum_i (p_i^z - 1)\tilde{v}_t(i).$$

Then, for such a state

$$\begin{aligned}
\tilde{v}_{t+1}(s) + \gamma \tilde{v}_{t,\Sigma} &= \gamma \left(D_t(s, s) \sum_i p_i^{z^*} \tilde{v}_t(i) + (1 - D_t(s, s)) \sum_i p_i^z \tilde{v}_t(i) \right) \\
&\leq -\delta + \gamma \sum_i (p_i^z - 1)\tilde{v}_t(i) + \gamma \tilde{v}_{t,\Sigma} \\
&\leq -\delta + \gamma \sum_i p_i^z \tilde{v}_t(i).
\end{aligned}$$

which implies

$$\gamma D_t(s, s) \sum_i (p_i^{z^*} - p_i^z)\tilde{v}_t(i) \leq -\delta.$$

We further have

$$-\text{sp}(\tilde{v}_t) = \min_i \tilde{v}_t(i) - \max_i \tilde{v}_t(i) \leq \sum_i \left(p_i^{z^*} - p_i^z \right) \tilde{v}_t(i) \leq -\frac{\delta}{\gamma D_t(s, s)} \leq -\frac{\delta}{D_t(s, s)}, \tag{10}$$

where the first inequality follows from $\max_i \tilde{v}_t(i) \geq \sum_i p_i^z \tilde{v}_t(i)$ and $\min_i \tilde{v}_t(i) \leq \sum_i p_i^{z^*} \tilde{v}_t(i)$.

Thus, for all states, (9) and (10) gives

$$D_t(s,s) \geq \min\left[\frac{\delta}{\text{sp}(\tilde{v}_t)}, 1\right] \geq \min\left[\frac{\delta}{\text{sp}(\tilde{v}_0)}, 1\right] = \delta',$$

where we set $\delta' = 1$ if $\text{sp}(\tilde{v}_0) = 0$.

This allows us to represent the update matrix P'_t as

$$P'_t = \tilde{P}'_t + \delta' P^*, \tag{11}$$

where

$$\tilde{P}'_t := (D_t - \delta' I)P^* + (I - D_t)P_t.$$

Since $D_t - \delta' I$ and $I - D_t$ are nonnegative diagonal matrices, $\tilde{P}'_t$ is nonnegative, and each of its rows sums to $1 - \delta'$. Hence, $\tilde{P}'_t$ is substochastic, and $\delta' P^*$ is a component that remains constant across iterations.

Step 4: Next, we analyze the value vector after $T = n^2$ iterations. Define

$$K_T := P'_T P'_{T-1} \cdots P'_1.$$

By Lemma D.1,

$$\tilde{v}_T = \gamma^T K_T \tilde{v}_0 - \gamma^T \tilde{E}\tilde{v}_0,$$

where $\tilde{E}$ has identical rows.

Define $\lambda_i^s := [K_T]_{si}$. Since K_T is stochastic,

$$\lambda_i^s \geq 0, \qquad \sum_i \lambda_i^s = 1.$$

Thus, λ_i^s is the weight with which $\tilde{v}_0(i)$ contributes to $\tilde{v}_T(s)$. Denoting the kth entry of each row of $\tilde{E}$ by $\tilde{E}_k$, we have

$$\tilde{v}_T(s) = \gamma^T \sum_i \lambda_i^s \tilde{v}_0(i) - \gamma^T \sum_k \tilde{E}_k \tilde{v}_0(k).$$

We now use the fact that the policy is unichain. Let us denote the set of states in the recurrent class as $\mathcal{S}_R$ and the set of transient states as $\mathcal{S}_T$. Then, we can rewrite the expression above as:

$$\tilde{v}_T(s) = \gamma^T \sum_{i \in \mathcal{S}_R} \lambda_i^s \tilde{v}_0(i) + \gamma^T \sum_{j \in \mathcal{S}_T} \lambda_j^s \tilde{v}_0(j) - \gamma^T \sum_k \tilde{E}_k \tilde{v}_0(k).$$

Step 5: It remains to lower bound the coefficients λ_i^s associated with recurrent states. Using (11),

$$K_T = P'_T \cdots P'_1 = \prod_{t=1}^{T} \left(\tilde{P}'_t + \delta' P^*\right).$$

All terms in this expansion are nonnegative. In particular, the term obtained by selecting $\delta' P^*$ from every factor is

$$(\delta')^T (P^*)^T.$$

Therefore, entrywise,

$$K_T \geq (\delta')^T (P^*)^T.$$

By the choice of T and the unichain and aperiodicity properties of P^*, for every $s \in \mathcal{S}$ and every $i \in \mathcal{S}_R$, the entry $[(P^*)^T]_{si}$ is positive. Let $P^*_{\min}$ denote the minimum nonzero entry of P^*. Then,

$$\lambda_i^s = [K_T]_{si} \geq (\delta')^T [(P^*)^T]_{si} \geq (\delta' P^*_{\min})^T =: \phi > 0, \qquad i \in \mathcal{S}_R.$$

Since K_T is stochastic, $\sum_i \lambda_i^s = 1$. Therefore,

$$|\mathcal{S}_R|\phi \leq \sum_{i \in \mathcal{S}_R} \lambda_i^s \leq 1.$$

Hence,

$$0 < |\mathcal{S}_R|\phi \leq 1.$$

Fix any $\epsilon \in (0, \phi)$ and define

$$\bar{\phi} := \phi - \epsilon \in (0, \phi).$$

Since $\lambda_i^s \geq \phi > \bar{\phi}$ for every $i \in \mathcal{S}_R$, we have

$$0 < |\mathcal{S}_R|\bar{\phi} < 1.$$

Let ν denote the uniform distribution over the recurrent class,

$$\nu(i) = \begin{cases} 1/|\mathcal{S}_R|, & i \in \mathcal{S}_R, \\ 0, & i \notin \mathcal{S}_R. \end{cases}$$

Defining

$$\alpha := |\mathcal{S}_R|\bar{\phi} \in (0, 1),$$

the bound $\lambda_i^s \geq \bar{\phi}$ implies that the previously defined T-step kernel K_T satisfies

$$K_T(s, \cdot) \geq \alpha\nu(\cdot), \qquad \forall s \in \mathcal{S}.$$

This is a Doeblin minorization. Equivalently, each row of K_T contains the same recurrent-state component of total weight α.

Step 6: We can then subtract $\bar{\phi}$ from each coefficient λ_i^s to obtain:

$$\tilde{v}_T(s) = \gamma^T \sum_{i \in \mathcal{S}_R} (\lambda_i^s - \bar{\phi})\tilde{v}_0(i) + \gamma^T \bar{\phi} \sum_{i \in \mathcal{S}_R} \tilde{v}_0(i) + \gamma^T \sum_{j \in \mathcal{S}_T} \lambda_j^s \tilde{v}_0(j) - \gamma^T \sum_k \tilde{E}_k \tilde{v}_0(k).$$

Having this expression, we can derive upper and lower bounds on the entries of $\tilde{v}_T(s)$ in terms of the maximum and minimum entries of $\tilde{v}_0(i)$. Combining the states back, we have:

$$\tilde{v}_T(s) \geq \gamma^T \left((1 - \bar{\phi}|\mathcal{S}_R|) \min_i \tilde{v}_0(i) + \bar{\phi} \sum_{i \in \mathcal{S}_R} \tilde{v}_0(i) - \sum_k \tilde{E}_k \tilde{v}_0(k) \right)$$

and

$$\tilde{v}_T(s) \leq \gamma^T \left((1 - \bar{\phi}|\mathcal{S}_R|) \max_i \tilde{v}_0(i) + \bar{\phi} \sum_{i \in \mathcal{S}_R} \tilde{v}_0(i) - \sum_k \tilde{E}_k \tilde{v}_0(k) \right).$$

Subtracting one from another gives the bound on the span of $\tilde{v}_T$:

$$\text{sp}(\tilde{v}_T) \leq \gamma^T (1 - \bar{\phi}|\mathcal{S}_R|)\text{sp}(\tilde{v}_0)$$

Denoting $\iota := 1 - \bar{\phi}|\mathcal{S}_R|$ completes the proof of the theorem. □

We now proceed to the proofs of both Corollaries, which start with the following lemma:

Lemma D.2. *For any SAP $z \in \pi_t$, where π_t is a policy implied by values v_t*

$$r^z \geq -\gamma \frac{sp(v_t)}{C}.$$

Proof. Denote the SAP chosen by the optimal policy in state s by z^*. Since SAP z is a greedy choice with respect to v_t, its advantage with respect to this value vector is larger, which yields:

$$0 \leq \text{adv}(z, v_t) - \text{adv}(z^*, v_t) = r^z - r^{z^*} + \sum_i (c_i^z - c_i^{z^*}) v_t(i)$$

$$\leq r^z + \gamma \frac{\text{sp}(v_t)}{C},$$

where the last inequality follows from the fact that optimal actions have 0 rewards in a normalized MDP and the redistribution argument applied to values $v_t(i)$ and coefficients $(c_i^z - c_i^{z^*})$. □

This lemma allows us to prove the corollaries.

Proof of Corollary 4.3. Run the algorithm for

$$t = \frac{\log(1/\epsilon) + \log(1/(1-\gamma)) + \log(\text{sp}(v_0)/C)}{\log(1/\gamma) + \log(1/\iota)/n^2}$$

iterations. By Theorem 4.2, after this number of iterations the obtained vector v_t satisfies

$$\frac{\text{sp}(v_t)}{C} \leq \epsilon(1-\gamma).$$

Applying Lemma D.2, we obtain that for all SAPs in π_t their rewards are lower bounded by $-\gamma\epsilon(1-\gamma)$, which in turn implies that the standard value function V^{π_t} is lower bounded by ϵ.

Since normalization preserves differences in values between policies, the same bound holds for the original MDP. □

The proof of the second corollary is almost identical to the first one.

Proof of Corollary 4.4. Run the algorithm for

$$t = \frac{\log(1/\epsilon) + \log(\text{sp}(v_0)/C)}{\log(1/\iota)/n^2}$$

iterations. By Theorem 4.2, after this number of iterations the obtained vector v_t satisfies

$$\frac{\text{sp}(v_t)}{C} \leq \epsilon.$$

Applying Lemma D.2, we obtain that for all SAPs in π_t their rewards are lower bounded by $-\epsilon$, which in turn implies that the gain of this policy $\rho^{\pi_t}(s)$ is lower bounded by ϵ for any state s.

Since normalization preserves differences in gains between policies (via the preservation of value differences and Lemma B.4), the same bound holds for the original MDP. □

E Additional Discussion

E.1 Algebraic and Geometric Intuition

Algebraic motivation. When $\gamma = 1$, the classical policy-evaluation matrix $I - P^\pi$ is singular because $P^\pi \mathbf{1} = \mathbf{1}$. Writing $\tilde{v}^\pi = v^\pi/C$, our representation instead solves

$$(I + \gamma E - \gamma P^\pi)\tilde{v}^\pi = R^\pi, \qquad E = \mathbf{1}\mathbf{1}^\top.$$

The additional term $\gamma E\tilde{v}^\pi$ is identical at every state and therefore leaves state differences unchanged while fixing the otherwise undetermined constant component. In particular, at $\gamma = 1$, multiplying by the stationary distribution of a unichain policy gives $\mathbf{1}^\top\tilde{v}^\pi = \rho^\pi$, so the same equation becomes $(I - P^\pi)\tilde{v}^\pi = R^\pi - \rho^\pi\mathbf{1}$. Thus, $\tilde{v}^\pi$ represents the bias with a specified additive constant. This provides a common algebraic representation for both reward criteria without requiring geometric arguments.

Geometric interpretation. The policy values determine a hyperplane in SAP space. Normalization makes the optimal-policy hyperplane horizontal: optimal SAPs have zero rewards, whereas non-optimal SAPs have negative rewards under the strict advantage gap. Consequently, the normalized iterates describe deviations from the optimal values, and $\text{sp}(v_t)$ measures their variation across states. Algebraically, this is an analysis of value errors; geometrically, it describes the flattening of the iterate hyperplanes toward the optimal hyperplane, up to a common value shift.

Why additional contraction occurs. The strict advantage gap guarantees that every stochastic update matrix contains a uniformly positive component of $P^\star$. Unichain structure and aperiodicity then ensure that, over a sufficiently long block, all rows of the update kernel share positive mass on the recurrent states. Writing this Doeblin minorization as

$$K_T = \alpha \mathbf{1}\nu^\top + (1-\alpha)Q, \qquad \alpha \in (0,1),$$

where ν is a probability vector and Q is stochastic, gives

$$\mathrm{sp}(K_T x) = (1-\alpha)\mathrm{sp}(Qx) \le (1-\alpha)\mathrm{sp}(x).$$

The common component cancels in the span, as do the state-independent shifts generated by E. Together with the discount factor, this yields the block contraction $\gamma^T(1-\alpha)$, with $1-\alpha = \iota$.

E.2 Conservativeness of the Bound

The bound in Theorem 4.2 is intentionally conservative. Its purpose is to establish the existence of an additional MDP-dependent geometric contraction, rather than to provide a tight prediction of the empirical convergence rate.

There are several sources of conservativeness. First, the proof uses the minimum nonzero transition probability $P^\star_{\min}$ together with the uniform block length $T = n^2$. This effectively tracks only a guaranteed minimal amount of probability mass that can propagate through the recurrent class over T steps. In a given MDP, the optimal-policy-induced chain may mix much faster, so substantially shorter blocks and larger effective weights are possible.

Second, we lower bound the diagonal entries of D_t using the initial span $\mathrm{sp}(\tilde{v}_0)$. In the actual dynamics, the relevant quantity depends on $\mathrm{sp}(\tilde{v}_t)$. As the span contracts, these diagonal weights increase, so the contribution of $P^\star$ becomes progressively larger. In fact, the entries $D_t[i,i]$ approach 1 as the iterates converge, which means that the update becomes increasingly dominated by the optimal-policy dynamics. The proof does not exploit this improvement and instead uses a single uniform lower bound valid for all iterations.

Third, the argument isolates only a small common portion of probability mass, namely the component of size $\bar{\phi}$ that is guaranteed to be present for every state. The remaining probability mass is treated only through worst-case bounds. In practice, this remaining mass can also be redistributed in a way that further reduces the span, but this additional contraction is not captured by the theorem. Thus, the factor

$$\iota = 1 - \bar{\phi}|\mathcal{S}_R|$$

should be interpreted as a worst-case contraction certificate rather than as an exact convergence rate.

Consequently, we do not claim that ι is tight or uniformly bounded away from 1 over the class of MDPs satisfying Assumption 4.1. The empirical rate can be substantially faster than the theoretical bound.

Figure 3 illustrates this point. It reproduces the experiments from Figure 1 and additionally overlays the theoretical bounds from Theorem 4.2. The observed convergence is considerably faster, consistent with the conservative nature of the estimates used in the proof.

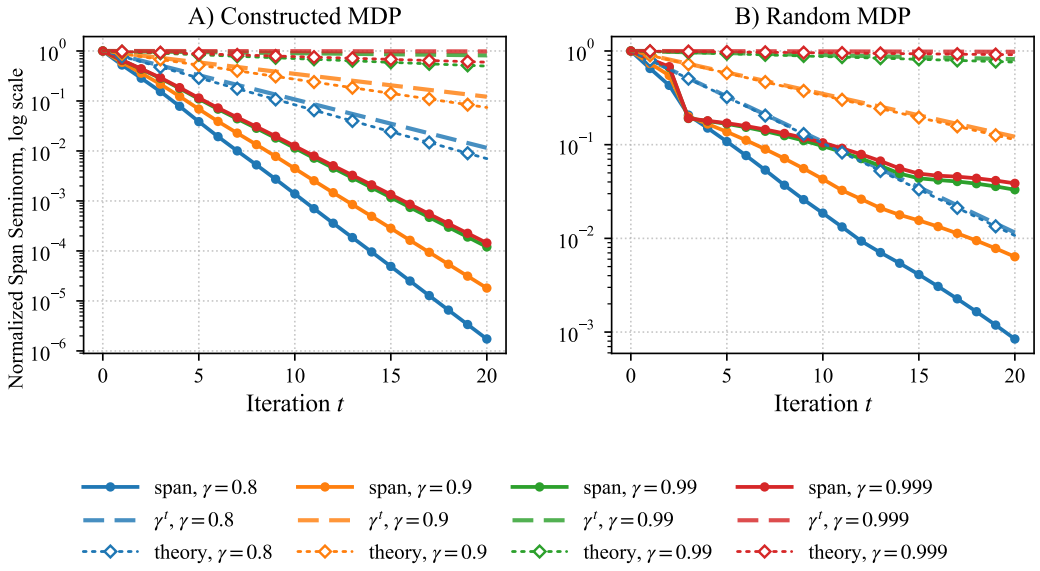


Figure 3: The same experiments as in Figure 1, with the theoretical convergence bounds from Theorem 4.2 added. The empirical convergence is substantially faster than the worst-case theoretical bound.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We believe that the claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations together with a validity of the assumptions in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Yes, the paper provides the full set of assumptions and proofs for the claimed theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification:

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is purely theoretical paper and we believe there is no societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.