Algorithms with Smoothed Polynomial-Time Complexity for Deterministic Discounted-sum and Mean-payoff Games

Ali Asadi
IST Austria
ali.asadi@ist.ac.at

Krishnendu Chatterjee
IST Austria
kchatterjee@ist.ac.at

Ruichen Luo
IST Austria
rluo@ist.ac.at

Abstract

Turn-based deterministic games on graphs are two-player zero-sum games played on finite directed graphs, where the vertex set is partitioned between the Max and the adversarial Min Player. The respective player chooses outgoing edges from their vertices. We consider these games with the classical discounted-sum and mean-payoff objectives. The goal of the Max Player is to maximize the objective against the adversarial Min Player. These games lie in NP $\cap$ coNP and are among the rare combinatorial problems that belong to this complexity class, yet the existence of a polynomial-time algorithm is a major open question. All known deterministic algorithms require exponential time in the worst case, while the classical strategy iteration (policy iteration) algorithm is remarkably efficient in practice, which motivates the study of the smoothed complexity of these games. Previous results established a smoothed polynomial-time bound only for the restricted class of ergodic games and presented only a high-probability guarantee. Our main result is a new analysis technique for a variant of the strategy iteration algorithm, and we show these algorithms achieve smoothed polynomial complexity for general game graphs with both high-probability and expected runtime guarantees. In particular, our result removes the ergodicity restriction of prior work and strengthens the high-probability guarantee to the stronger expected-runtime guarantee, thereby resolving the open question on the smoothed complexity of these games.[1]

[1]Through personal communication, we are aware of independent and concurrent work by Bruno Loff and Mateusz Skomra, titled "Smoothed analysis of policy iteration and value iteration for deterministic discounted and mean-payoff games played on any graph".

1 Introduction

Turn-based game graphs. The class of two-player turn-based (perfect-information) deterministic games is a widely studied theoretical model. The model consists of a finite directed graph $G = (V, E)$ where the vertices are partitioned into Player-Max and Player-Min vertices, and the corresponding player chooses outgoing edges of the vertices in control. This model extends classical graphs and is used to model interaction between two adversarial agents, and has many applications, e.g., in analysis of reactive synthesis (interaction of controller and environment) [PR89, RW87]; in adversarial planning [MB85, HZ98]. Moreover, this model closely corresponds to the notion of alternation in Turing machines which has been studied in [CKS81].

Discounted-sum and mean-payoff objectives. In the study of games over graphs, the two fundamental payoff functions or objectives are discounted-sum and mean-payoff. Every edge is assigned a real-valued reward, which represents the per-stage reward. A play (or infinite-walk) of the graph is an infinite-sequence of vertices, with edges between the consecutive vertices. The payoff of a play under discounted-sum (resp., mean-payoff) is the discounted-sum (resp., the long-run average) of the rewards that appear in the play. These objectives are the most well-studied and basic objectives studied in games over graphs [Sha53, Gil57, FV96, Put94], e.g., mean-payoff and discounted-sum games have been studied for quantitative reactive synthesis [DAHM03, CPV15, BCHJ09].

Algorithmic results. Besides the practical motivation to study turn-based deterministic games with discounted-sum and mean-payoff objectives, they represent an intriguing class of problems for algorithmic study. The main computational problem is related to the optimal value for an objective, where the Max Player aims to maximize and the Min Player aims to minimize the payoff function. The decision problems for these games lie in NP $\cap$ coNP (also UP $\cap$ coUP) [Con92, ZP96], but the existence of a polynomial-time algorithm is a long-standing and major open problem. All deterministic algorithms for this problem have a worst-case exponential-time complexity.

Smoothed complexity. While the worst-case complexity for the computational problem for these games is exponential, a classical algorithm (namely the policy-iteration algorithm) works very well in practice. Hence it is natural to consider the smoothed complexity of this problem [ST04], which is a classical complexity analysis between worst-case and average-case analysis. The smoothed complexity analysis of these problems has an interesting history: First, a polynomial-time smoothed complexity result for mean-payoff games was claimed in [BEF$^+$11]; unfortunately this claim was incorrect and later retracted. Second, this result was partially recovered in [LS24] which shows smoothed polynomial-time complexity for the special class of ergodic games (where all vertices have the same value) with a high-probability guarantee.

Open problem. The class of ergodic games is often simpler to analyze, e.g., partially-observable MDPs (POMDPs) with ergodicity are decidable [CLSZ25] whereas in general they are undecidable [MHC03]; and for mean-payoff objectives the algorithmic study for ergodic games is much simpler [CIJ14, HK66]. Hence the first key open problem is to extend the smoothed polynomial time result from ergodic to general game graphs. Moreover, the previous result only presents a high-probability guarantee, whereas for smoothed complexity analysis a stronger and more desired guarantee is the expected runtime guarantee (see Remark 1). Thus, the second interesting open problem is to establish polynomial-time smoothed complexity with expected runtime guarantee for turn-based deterministic discounted-sum and mean-payoff games.

Our contribution. We answer the above open questions in the affirmative. Moreover, we show that variants of the policy-iteration algorithm achieve the polynomial-time smoothed complexity on general game graphs with expected runtime guarantee. In other words, we resolve the open question with variants of a well-known algorithm that also works well in practice.

Related works. The algorithmic study of discounted-sum and mean-payoff games has a long history [EM79, GKK88, ZP96, ACSSU24, DKZ19, Koz21]. All these deterministic algorithms have worst-case exponential-time complexity for the general case, and even the known randomized algorithms achieve only sub-exponential time [Lud95, BV07]. Moreover, discounted-sum games are closely related to turn-based stochastic games, which subsume MDPs as a special case. Thus, it is natural to ask whether our smoothed polynomial-time result extends to this broader class. Quite interestingly, [CY23] establishes a super-polynomial-time lower bound for the smoothed complexity of policy iteration on MDPs, which rules out such an extension to turn-based stochastic games.

2 Preliminaries

For ease of reference, Table 1 in Appendix A collects the recurring notation of the paper.

Definition 1 (Game graphs)**.** A game graph $G = (V_{\min}, V_{\max}, E)$ is a directed finite graph consisting of two disjoint finite sets of vertices $V_{\min}$ and $V_{\max}$, with vertex set $V \triangleq V_{\min} \uplus V_{\max}$ and edges $E \subseteq V \times V$. We denote by $n \triangleq |V|$ and $m \triangleq |E|$. An edge $a = (v, v') \in E$ indicates the transition from vertex v to v'. Without loss of generality, we consider that for every vertex $v \in V$, there exists at least one $v' \in V$ such that $(v, v') \in E$.

Dynamics and paths. At the beginning of the game, a token is placed at an initial vertex $v_0 \in V$. At any time t ($t \in \mathbb{Z}_{\geq 0}$): if $v_t \in V_{\min}$, then the Min Player chooses an edge $a_t \in E$, where $a_t = (v_t, v_{t+1})$, and the token is moved from v_t to v_{t+1}; and similarly, if $v_t \in V_{\max}$, then the Max Player chooses an edge and moves the token. A *path* (or play) starting from an initial vertex v_0 is an infinite sequence of vertices $\rho = v_0, v_1, v_2, \ldots$ such that $(v_t, v_{t+1}) \in E$ for all $t \in \mathbb{Z}_{\geq 0}$.

Objectives. We consider *discounted-sum* and *mean-payoff* objectives, and in both cases there is a reward vector $\mathbf{r} \in \mathbb{R}^E$ that assigns to each edge $a = (v, v') \in E$ a real number $\mathbf{r}(v, v')$ as its per-stage reward.

- *Discounted-sum objective:* For the discounted-sum objective, a real discount factor $\gamma \in [0, 1)$ is given, and the game is denoted by $G^{\gamma,\mathbf{r}}$. For any path $\rho = v_0, v_1, v_2, \ldots$, the discounted-sum payoff is the infinite discounted sum of the rewards of the edges that appear in the path:

$$u^{\gamma,\mathbf{r}}(\rho) = \sum_{t=0}^{\infty} \gamma^t \mathbf{r}(v_t, v_{t+1}).$$

  Since $\gamma \in [0, 1)$ and the graph is finite (i.e., rewards are bounded), this infinite sum converges.
- *Mean-payoff objectives.* For mean-payoff objectives, the same graph and reward vector are considered, denoted by $G^{\mathrm{mp},\mathbf{r}}$, but the payoff function evaluates a path by its long-run average reward instead of discounted sum. For any path $\rho = v_0, v_1, \ldots$, the *mean-payoff* is:

$$u^{\mathrm{mp},\mathbf{r}}(\rho) = \liminf_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{r}(v_t, v_{t+1}).$$

Strategies. A *strategy* (or policy) is a rule that a player uses to choose their next edge. In the most general sense, a strategy can depend on the prefixes of paths. However, it is known that for both discounted-sum and mean-payoff games with perfect information, memoryless and deterministic optimal strategies always exist [Con92, EM79]. We refer to strategies that are both memoryless and deterministic as *positional* strategies, to which we restrict our attention without loss of generality. Formally, we define a strategy for the Min Player as a mapping $\pi_{\min} : V_{\min} \to V$ such that $(v, \pi_{\min}(v)) \in E$ for all $v \in V_{\min}$. Symmetrically, a strategy for the Max Player is a mapping $\pi_{\max} : V_{\max} \to V$ such that $(v, \pi_{\max}(v)) \in E$ for all $v \in V_{\max}$. We denote the sets of all valid positional strategies for the Min and Max Players as $\Pi_{\min}$ and $\Pi_{\max}$, respectively. Given an

initial vertex $v \in V$ and a pair of positional strategies $(\pi_{\min}, \pi_{\max})$, the game graph G produces a unique path, which we denote by $\rho(v, \pi_{\min}, \pi_{\max})$. Note that this path consists of a finite-path followed by a simple cycle infinitely repeated.

Values. It is well-known from the literature [Con92] that for any given γ and $\mathbf{r}$, every vertex $v \in V$ has a unique discounted-sum *value*, denoted by $\mu^{\gamma,\mathbf{r}}(v)$, which satisfies:

$$\mu^{\gamma,\mathbf{r}}(v) = \min_{\pi_{\min} \in \Pi_{\min}} \max_{\pi_{\max} \in \Pi_{\max}} u^{\gamma,\mathbf{r}}(\rho(v, \pi_{\min}, \pi_{\max})) = \max_{\pi_{\max} \in \Pi_{\max}} \min_{\pi_{\min} \in \Pi_{\min}} u^{\gamma,\mathbf{r}}(\rho(v, \pi_{\min}, \pi_{\max})).$$

Similarly, the mean-payoff game $G^{\mathrm{mp},\mathbf{r}}$ has a unique value for every $v \in V$, denoted by $\mu^{\mathrm{mp},\mathbf{r}}(v)$.

Optimal strategies. A strategy $\pi^*_{\min} \in \Pi_{\min}$ is an *optimal min strategy* for the discounted-sum game $G^{\gamma,\mathbf{r}}$ if it guarantees that the discounted-sum payoff never exceeds the discounted-sum value, regardless of the opponent's strategy; i.e., for all starting vertices $v \in V$ and all $\pi_{\max} \in \Pi_{\max}$, we have $u^{\gamma,\mathbf{r}}(\rho(v, \pi^*_{\min}, \pi_{\max})) \leq \mu^{\gamma,\mathbf{r}}(v)$. The notion of optimal max strategies is defined symmetrically. We say that $(\pi^*_{\min}, \pi^*_{\max})$ is a pair of optimal strategies for $G^{\gamma,\mathbf{r}}$ if $\pi^*_{\min}$ is an optimal min strategy and $\pi^*_{\max}$ is an optimal max strategy for the discounted-sum objective. The notion of optimal strategies for mean-payoff games is defined analogously.

3 Smoothed model

The current best-known deterministic algorithms for discounted-sum and mean-payoff games have worst-case exponential complexity. However, in practice, instances of these games rarely exhibit worst-case behaviors. Hence the goal of this work is to study the problems through the lens of *smoothed analysis* [ST04], which bridges the gap between average-case and worst-case complexity by analyzing the performance of an algorithm under slight random perturbations of adversarial inputs. Throughout this section, we describe the smoothed model for discounted-sum games. The model for mean-payoff games is obtained similarly.

Input formalization. We follow the standard notion of smoothed model where the adversarial input is normalized. In discounted-sum and mean-payoff games, the rewards can be shifted and scaled to ensure that they lie in the interval $[-1, 1]$. In our smoothed analysis framework, an adversary first specifies the graph $G = (V_{\min}, V_{\max}, E)$ and the discount factor $\gamma \in (0, 1)$. The adversary also chooses an initial, worst-case reward vector $\mathbf{r}_0 \in [-1, 1]^E$. That is, the adversary can assign any initial reward to any edge, provided it is bounded within the interval $[-1, 1]$. After the adversary fixes G, γ, and $\mathbf{r}_0$, a random continuous perturbation is applied to the rewards. Specifically, we define a noise vector $\xi \in \mathbb{R}^E$. For broad application, we do not restrict ξ to a specific distribution. Instead, we require the perturbation to satisfy a set of standard regularity properties, parameterized by a maximum density bound $\phi > 0$ and a tail decay parameter $\theta > 0$.

Definition 2 (Smoothed model)**.** The random noise vector $\xi \in \mathbb{R}^E$ satisfies the following properties:

1. *Independence:* The coordinates ξ_a for all $a \in E$ are mutually independent.
2. *Density bound ϕ:* For all $a \in E$, the probability distribution of ξ_a is absolutely continuous with respect to the Lebesgue measure. Furthermore, its probability density function is uniformly bounded by ϕ.
3. *Tail bound θ:* For all $a \in E$ and $t > 0$, the tail probability satisfies $\mathbb{P}[|\xi_a| > t] \leq \exp(-\theta\, t)$.

Special cases: Our general smoothed model captures many standard continuous distributions:

- *Gaussian noise:* For $\xi_a \sim \mathcal{N}(0, \sigma^2)$, the conditions hold with $\phi = \frac{1}{\sigma\sqrt{2\pi}}$ and $\theta = \frac{\sqrt{2/\pi}}{\sigma}$.
- *Laplacian noise:* For $\xi_a \sim \mathrm{Laplace}(0, b)$, the conditions hold with $\phi = \frac{1}{2b}$ and $\theta = \frac{1}{b}$.
- *Uniform noise:* For $\xi_a \sim \mathrm{Uniform}(-b, b)$, the conditions hold with $\phi = \frac{1}{2b}$ and $\theta = \frac{1}{b}$.

Problem statement. Given the input consisting of a game graph G, discount factor γ, and the normalized reward vector $\mathbf{r}_0$, the algorithmic problem is to compute optimal strategies and values for the *smoothed game* $G^{\gamma,\mathbf{r}}$, where the final perturbed reward vector $\mathbf{r} \in \mathbb{R}^E$ is given by $\mathbf{r} = \mathbf{r}_0 + \xi$.

Probability measure. The random noise ξ induces a unique probability measure over the space of perturbed reward vectors $\mathbf{r}$. We define the probability measure as $\mathbb{P}_\xi[\cdot]$ and the corresponding expectation under this measure as $\mathbb{E}_\xi[\cdot]$. Since for any edge a the adversarial base rewards are bounded such that $\mathbf{r}_0(a) \in [-1, 1]$, the tail concentration property guarantees that the magnitude of the final perturbed reward satisfies

$$\mathbb{P}_\xi[|\mathbf{r}(a)| > 1 + t] \leq \exp\big(-\theta\, t\big). \tag{1}$$

Polynomial-time smoothed complexity. Finally, we define the smoothed complexity of an algorithm for this problem. Let $\mathcal{T}(G, \gamma, \mathbf{r}_0 + \xi)$ denote the total number of arithmetic operations of an algorithm on the perturbed game $G^{\gamma,\mathbf{r}}$. We evaluate the algorithm's performance under two standard types of smoothed guarantees:

1. *High probability guarantee:* An algorithm achieves *smoothed polynomial time with high probability* if, for any failure probability $\epsilon \in (0, 1)$, its runtime is bounded by a polynomial in the game dimensions (n and m), the noise parameters (ϕ and $1/\theta$), and the inverse failure probability ($1/\epsilon$), with probability at least $1 - \epsilon$. Formally, for any graph G, discount factor γ, initial reward vector $\mathbf{r}_0 \in [-1, 1]^E$:
$$\mathbb{P}_\xi\Big[\mathcal{T}(G, \gamma, \mathbf{r}_0 + \xi) \leq \mathrm{poly}\Big(n, m, \phi, \frac{1}{\theta}, \frac{1}{\epsilon}\Big)\Big] \geq 1 - \epsilon.$$
2. *Expected runtime guarantee:* An algorithm achieves *expected smoothed polynomial time* if its expected runtime over the random noise is bounded by a polynomial in n, m, ϕ, and $1/\theta$. Formally, for any graph G, discount factor γ, and initial reward vector $\mathbf{r}_0 \in [-1, 1]^E$:
$$\mathbb{E}_\xi\Big[\mathcal{T}(G, \gamma, \mathbf{r}_0 + \xi)\Big] \leq \mathrm{poly}\Big(n, m, \phi, \frac{1}{\theta}\Big).$$

Remark 1 (High probability vs. expectation guarantee). We clarify the two types of guarantees above in the context of smoothed complexity. In the design of randomized algorithms, a high-probability bound is often considered a stronger guarantee because the algorithm can draw fresh randomness and restart upon failure. In smoothed analysis, however, the random perturbation is fixed exactly once, making restarts impossible. Consequently, an expected polynomial runtime is a strictly stronger mathematical guarantee. As specifically noted in [ST09], a high-probability polynomial bound follows from an expected polynomial bound via Markov's inequality. The converse, however, does not hold: an algorithm may run efficiently on $1 - \epsilon$ fraction of the perturbations, yet its runtime could grow exponentially on the remaining ϵ fraction of bad instances, resulting in exponential expected runtime. Thus, bounding the expected runtime is the desired and stronger guarantee in the context of smoothed analysis. Thus the main goal of the study is to provide above guarantees, especially expected runtime guarantee, for discounted-sum and mean-payoff games.

4 Overview of Techniques

Previous result and technique. The previous result [LS24] in the literature for polynomial-time smoothed complexity for games with mean-payoff and discounted-sum objectives is restrictive in two ways: First, the class of game graphs is restricted to ergodic game graphs, where all vertices have the same value for mean-payoff objectives. Second, the result only provides smoothed high-probability guarantee and not the more general expected guarantee. The main algorithm for the smoothed analysis is the policy-iteration algorithm, which is simpler for ergodic mean-payoff games,

as compared to general mean-payoff games. While for general mean-payoff games, the policy-iteration algorithm for mean-payoff objectives relies on (i) values and (ii) potential or bias; for ergodic games, since all vertices have the same value, the only relevant concept is the bias. The results of [LS24] depend on the notion of bias-induced optimal strategies, which is only relevant for ergodic games. The key mathematical techniques in [LS24] along with bias-induced optimal strategies are (i) the notion of condition number, and (ii) geometric interpretation of reward vectors and hyper-plane separation induced by strategy profiles.

Our techniques: Our analysis provides the general result removing all the above restrictions. We present our key technical contributions below.

1. *Ingredient 1:* First, we provide a quantitative decomposition of discounted-sum values with respect to mean-payoff values (Lemma 2) and then present a Lipschitz continuity result with respect to switches of edges from optimal strategies (Lemma 3). These results hold for all game graphs, and are neither specific to the smoothed model nor a probabilistic guarantee.
2. *Ingredient 2:* Second, we present a probabilistic separation between (i) the value of a vertex and (ii) the value obtained by playing a sub-optimal edge and then playing optimally (Lemma 4); which we call the separation of the sub-optimality gap.
3. *Ingredient 3:* Finally, for general mean-payoff games, the policy iteration algorithm depends on both the values and the bias. Our final key ingredient is a mean-payoff value separation bound, which shows that for every vertex pair either the value is the same or there is a separation bound for the values (Lemma 5).

Ingredient 1 is the foundational result which is the basis of all results. We remark that with Ingredient 1 and Ingredient 2 we have an alternative proof for ergodic games with high-probability guarantee. However, for brevity, we focus on general game graphs for this paper. Ingredient 3 along with the other two ingredients is the main idea for generalization from ergodic to all game graphs, for high-probability guarantee. For the extension from high-probability to expected runtime guarantee we present another technical contribution:

4. *Ingredient 4:* We consider the Blackwell-optimal discount factor γ_{bw} such that optimal strategies for discounted-sum are also optimal for the mean-payoff objectives. Let the inverse gap be $1/(1-\gamma_{\mathrm{bw}})$. The runtime complexity of the policy-iteration algorithm is dominated by the inverse gap, and in general the expected value of this inverse gap may be exponential. We therefore consider a truncated version of this inverse gap, which is the minimum of the inverse gap and the number of strategies, and show a polynomial upper bound for the expectation of the truncated version (Lemma 9).

Our ingredients tightly relate discounted-sum and mean-payoff games and hence we view mean-payoff games through the lens of discounted-sum games and Blackwell-optimality. Hence our main analysis focuses on discounted-sum games with arbitrary discount factors. We start with the formal background on Blackwell-optimality and policy iteration algorithm for discounted-sum games.

5 Background on Blackwell Optimality and Policy Iteration

In this section, we provide some useful theoretical and algorithmic tools from the literature. We first recall the concept of Blackwell optimality, which is a fundamental result in the theory of dynamic programming and games that establishes the existence of strategies, namely *Blackwell optimal strategies*, that are uniformly optimal for all discount factors sufficiently close to 1 [Bla62, BK76]. This property provides an important bridge between discounted-sum and mean-payoff games. We then describe the classic policy iteration algorithm for discounted-sum games to compute the optimal strategies and game values.

Blackwell threshold $\gamma_{\mathbf{bw}}(\mathbf{r})$ and Blackwell-optimal strategies. Formally, given a specified

reward vector $\mathbf{r}$, a strategy $\pi^{\mathrm{bw}}_{\min} \in \Pi_{\min}$ is a *Blackwell optimal min strategy* if there exists a threshold discount factor $\gamma_0 \in [0,1)$ such that for all $\gamma \in [\gamma_0, 1)$, the strategy $\pi^{\mathrm{bw}}_{\min}$ is an optimal min strategy for the discounted-sum game $G^{\gamma,\mathbf{r}}$. Symmetrically, a strategy $\pi^{\mathrm{bw}}_{\max} \in \Pi_{\max}$ is a *Blackwell optimal max strategy* with respect to $\mathbf{r}$ if there exists a threshold discount factor $\gamma_0 \in [0,1)$ such that for all $\gamma \in [\gamma_0, 1)$, $\pi^{\mathrm{bw}}_{\max}$ is an optimal max strategy for $G^{\gamma,\mathbf{r}}$. We define the *Blackwell threshold* with respect to $\mathbf{r}$, denoted by $\gamma_{\mathrm{bw}}(\mathbf{r}) \in [0,1)$, as the infimum over all such valid thresholds.

Asymptotic relationship between discounted-sum and mean-payoff values. By definition, for any $\gamma \in [\gamma_{\mathrm{bw}}(\mathbf{r}), 1)$, a pair of Blackwell optimal strategies $(\pi^{\mathrm{bw}}_{\min}, \pi^{\mathrm{bw}}_{\max})$ is guaranteed to be optimal for the discounted-sum game $G^{\gamma,\mathbf{r}}$. Furthermore, these strategies are of particular theoretical importance because they simultaneously act as optimal strategies for the corresponding mean-payoff game $G^{\mathrm{mp},\mathbf{r}}$, providing a crucial bridge between discounted-sum and mean-payoff games [BK76]. Formally, for any reward vector $\mathbf{r} \in \mathbb{R}^E$ and vertex $v \in V$, the normalized discounted-sum value converges exactly to the mean-payoff value, i.e.,

$$\mu^{\mathrm{mp},\mathbf{r}}(v) = \lim_{\gamma \uparrow 1} (1-\gamma)\mu^{\gamma,\mathbf{r}}(v). \tag{2}$$

Policy iteration. To compute the optimal strategies and values in the discounted-sum game $G^{\gamma,\mathbf{r}}$, we describe the classic policy iteration algorithm, formally presented in Algorithm 1 in Appendix B. The algorithm operates from the perspective of one of the players; without loss of generality, we formulate it for the Min Player. Starting from an arbitrary initial strategy, the algorithm iteratively refines the strategy $\pi_{\min}$ through alternating phases of policy evaluation and policy improvement.

- *Policy evaluation phase:* The algorithm fixes the Min Player's current strategy $\pi_{\min}$. With $\pi_{\min}$ fixed, the game reduces to a graph with only choices for the Max Player (aka deterministic MDP). Graphs (or deterministic MDPs) with discounted-sum objectives can be solved in strongly polynomial time [PY15]. The resulting vector μ represents the expected discounted-sum payoff under the current strategies.
- *Policy improvement phase:* The algorithm searches for *improving switches*—local edge changes that strictly decrease the current evaluated payoff for the Min Player. If no improving switch exists for any vertex in $V_{\min}$, the algorithm terminates. Otherwise, the algorithm updates $\pi_{\min}$ by simultaneously applying the best improving switch at every possible vertex. This specific greedy update mechanism is known as Howard's all-switch rule.

In other words, the policy-iteration algorithm iterates over strategies of a player by greedily switching over local improvements.

Correctness and complexity. The key correctness proof establishes that when no local-improvement switches are possible, then the obtained strategy is globally optimal [Con92]. Since the total number of strategies is at most n^n and each update strictly decreases the value vector, the algorithm is guaranteed to converge in finite time [FV96]. Specifically, for discounted-sum games, the upper bound on the number of iterations depends on the discount factor γ. The worst-case strongly polynomial bound for a constant discount factor from the literature, which is the basic result we use in subsequent analysis, is as follows:

Proposition 1 (HMZ13, Theorem 7.5)**.** *Let $\gamma \in [0,1)$. For any game $G^{\gamma,\mathbf{r}}$ with n vertices and m edges, the number of iterations that* PI *algorithm (Algorithm 1) requires is at most* $6\frac{m}{1-\gamma}\ln\left(\frac{n}{1-\gamma}\right)$.

Each iteration of PI performs a single policy evaluation, which requires at most $\mathcal{O}(n^3 m^2 \log^2 n)$ arithmetic operations [PY15]. While this algorithm performs well in practice, the worst-case bound of Proposition 1 is exponential as γ approaches 1 (e.g., $\gamma = 1 - 2^{-n}$).

6 Mathematical Ingredients 1-3

In this section, we present the details related to the first three ingredients mentioned in Section 4; full proofs are provided in Appendix C.

6.1 Ingredient 1: Value Decomposition and Lipschitz Continuity

In this subsection, we present conceptual ideas behind Ingredient 1 from Section 4.

- We first establish a value decomposition for discounted-sum games under Blackwell optimality (Lemma 2), which provides a useful connection between discounted-sum and mean-payoff games. As compared to Equation (2) which only provides convergence-in-the-limit guarantee, this result provides a quantitative bound on the difference of the discounted-sum and mean-payoff values. While a general characterization of discounted-sum and mean-payoff value is provided in [BK76] for concurrent games (with simultaneous interaction), we present a quantitative characterization albeit for the simpler class of turn-based games.
- We then define the suboptimality gap, which quantifies the suboptimality incurred by switching a single edge in the optimal strategy. We show that whenever the two endpoints of an edge share the same mean-payoff value, the suboptimality gap is Lipschitz continuous with respect to the discount factor γ over the Blackwell threshold (Lemma 3)

Lemma 2 (Value decomposition under Blackwell optimality)**.** *Consider a game graph G with reward vector $\mathbf{r} \in \mathbb{R}^E$. Let $v \in V$ be a vertex.*

1. Decomposition: *For all $\gamma \in (\gamma_{bw}(\mathbf{r}), 1)$, the discounted-sum value admits the following decomposition, where $w_v^{\mathbf{r}}(\gamma)$ is a rational function of γ:*
$$\mu^{\gamma,\mathbf{r}}(v) = \frac{\mu^{mp,\mathbf{r}}(v)}{1-\gamma} + w_v^{\mathbf{r}}(\gamma).$$
2. Magnitude and derivative: *For all $\gamma \in (\gamma_{bw}(\mathbf{r}), 1)$, the magnitude and the derivative of $w_v^{\mathbf{r}}(\gamma)$ are bounded by:*
$$|w_v^{\mathbf{r}}(\gamma)| \leq 4n\|\mathbf{r}\|_\infty \quad \textit{and} \quad \left|(\frac{d}{d\gamma} w_v^{\mathbf{r}})(\gamma)\right| \leq 7n^2\|\mathbf{r}\|_\infty.$$

Proof sketch. Let $\mu \triangleq \mu^{\mathrm{mp},\mathbf{r}}(v)$. Fix a Blackwell optimal strategy pair, under which the path from v consists of a transient path of length $t \leq n$ followed by a cycle of length $c \leq n$. Let r_k denote the reward at step k. Then μ is the average cycle reward, so $\sum_{j=0}^{c-1}(r_{t+j} - \mu) = 0$. Writing each reward as $(r_k - \mu) + \mu$ in $\mu^{\gamma,\mathbf{r}}(v) = \sum_{j=0}^{t-1} r_j\gamma^j + \frac{\gamma^t}{1-\gamma^c}\sum_{j=0}^{c-1} r_{t+j}\gamma^j$ shows that the μ terms contribute $\mu/(1-\gamma)$. The cycle cancellation removes this singular factor from the remaining terms, which define the rational function $w_v^{\mathbf{r}}(\gamma)$. For magnitude and derivative, $w_v^{\mathbf{r}}(\gamma)$ contains at most $2n$ deviations, each bounded by $2\|\mathbf{r}\|_\infty$ and multiplied by a coefficient in $[0,1]$, giving $|w_v^{\mathbf{r}}(\gamma)| \leq 4n\|\mathbf{r}\|_\infty$. Differentiating these coefficients bounds the transient contribution by $n^2\|\mathbf{r}\|_\infty$ and the cycle contribution by $6n^2\|\mathbf{r}\|_\infty$, yielding the derivative bound. □

Suboptimality gap $\Delta^{\gamma,\mathbf{r}}$. We define the *suboptimality gap* for the discounted-sum game $G^{\gamma,\mathbf{r}}$ as follows. For all edges $(v,v') \in E$,
$$\Delta^{\gamma,\mathbf{r}}(v,v') = \mathbf{r}(v,v') + \gamma\mu^{\gamma,\mathbf{r}}(v') - \mu^{\gamma,\mathbf{r}}(v);$$
i.e., the quantity $\mathbf{r}(v,v')+\gamma\mu^{\gamma,\mathbf{r}}(v')$ is the value for switching via the edge, and $\mu^{\gamma,\mathbf{r}}(v)$ is the optimal value; and hence the function represents the relative value for switching via the edge.

Main property. For any reward vector $\mathbf{r} \in \mathbb{R}^E$ such that $\mu^{\mathrm{mp},\mathbf{r}}(v) = \mu^{\mathrm{mp},\mathbf{r}}(v')$, we show that the suboptimality gap is Lipschitz continuous in γ over the Blackwell threshold. In contrast, the value vector $\mu^{\gamma,\mathbf{r}}(v)$ may not be Lipschitz continuous in γ because the singular term $\frac{\mu^{\mathrm{mp},\mathbf{r}}(v)}{1-\gamma}$ in the value decomposition (Lemma 2) diverges as γ approaches 1.

Lemma 3 (Lipschitz continuity)**.** *Let $\mathbf{r} \in \mathbb{R}^E$ be any reward vector. For any edge $a = (v, v') \in E$ where $\mu^{mp,\mathbf{r}}(v) = \mu^{mp,\mathbf{r}}(v')$, the map $\gamma \mapsto \Delta^{\gamma,\mathbf{r}}(v, v')$ is $18n^2\|\mathbf{r}\|_\infty$-Lipschitz continuous over the interval $(\gamma_{bw}(\mathbf{r}), 1)$.*

Proof sketch. Substituting the value decomposition from Lemma 2 into the suboptimality gap, the shared mean-payoff value cancels the singular terms, giving $\Delta^{\gamma,\mathbf{r}}(v, v') = \mathbf{r}(v, v') - \mu^{\mathrm{mp},\mathbf{r}}(v) + \gamma w^{\mathbf{r}}_{v'}(\gamma) - w^{\mathbf{r}}_v(\gamma)$. Therefore, using the magnitude and derivative bounds from Lemma 2, we have $\left|\frac{d}{d\gamma}\Delta^{\gamma,\mathbf{r}}(v, v')\right| \le 4n\|\mathbf{r}\|_\infty + 14n^2\|\mathbf{r}\|_\infty \le 18n^2\|\mathbf{r}\|_\infty$, which proves the result. □

Gap at one $\Delta^{1,\mathbf{r}}$. The Lipschitz continuity result implies that we can extend the definition of $\Delta^{\gamma,\mathbf{r}}$ to $\gamma = 1$ as follows. For any edge $(v, v') \in E$ where $\mu^{\mathrm{mp},\mathbf{r}}(v) = \mu^{\mathrm{mp},\mathbf{r}}(v')$, we have

$$\Delta^{1,\mathbf{r}}(v, v') \triangleq \lim_{\gamma \uparrow 1} \Delta^{\gamma,\mathbf{r}}(v, v')\,.$$

Following from Lemma 3, the limit exists and is finite. We also refer to $\Delta^{1,\mathbf{r}}$ as the *suboptimality gap* of the mean-payoff game $G^{\mathrm{mp},\mathbf{r}}$.

6.2 Ingredient 2: Suboptimality Gap Separation

We now establish a probabilistic separation result for the suboptimality gap under the smoothed model (Ingredient 2 from Section 4). For any fixed edge $a = (v, v')$, the event that (i) vertices v and v' have the same mean-payoff value and (ii) the suboptimality gap at $\gamma = 1$ is small and non-zero, has low probability.

Lemma 4 (Suboptimality separation)**.** *Let G be any game graph and let $a = (v, v') \in E$ be any fixed edge. Under the smoothed model, for any $\delta > 0$, we have*

$$\mathbb{P}_\xi\left[\mu^{mp,\mathbf{r}}(v) = \mu^{mp,\mathbf{r}}(v') \text{ and } 0 < \left|\Delta^{1,\mathbf{r}}(a)\right| < \delta\right] \le 2\delta\phi.$$

Proof sketch. Consider the event $\mathcal{E} \triangleq \left\{\mathbf{r} \in \mathbb{R}^E : \mu^{\mathrm{mp},\mathbf{r}}(v) = \mu^{\mathrm{mp},\mathbf{r}}(v') \text{ and } 0 < \left|\Delta^{1,\mathbf{r}}(a)\right| < \delta\right\}$. If a is the only outgoing edge from v, then $\Delta^{1,\mathbf{r}}(a) = 0$ whenever well-defined, so $\mathcal{E}$ is empty. Otherwise, fix the rewards $\mathbf{r}_{-a}$ on all edges except a and write $x \triangleq \mathbf{r}(a)$. For any x such that $(x; \mathbf{r}_{-a}) \in \mathcal{E}$, the edge a is strictly suboptimal for all γ sufficiently close to 1. Deleting a therefore leaves the value vector unchanged and preserves the shared mean-payoff value of v and v'. Thus,

$$\Delta^{1,\mathbf{r}}(a) = x + \lim_{\gamma \uparrow 1}\left(\gamma\mu^{\gamma,\mathbf{r}_{-a}}(v') - \mu^{\gamma,\mathbf{r}_{-a}}(v)\right).$$

By Lemma 2, the limit term is finite and independent of x. Hence the set of x values producing $\mathcal{E}$ has length at most 2δ, and therefore conditional probability at most $2\delta\phi$. Since this bound holds for every fixed $\mathbf{r}_{-a}$, the unconditional bound follows. □

6.3 Ingredient 3: Mean-payoff Value Separation

We establish Ingredient 3 from Section 4: under the smoothed model, the mean-payoff values of any two vertices are either the same or separated by an inverse-polynomial gap with high probability.

Lemma 5 (Mean-payoff separation)**.** *Let G be any game graph and let $v, v' \in V$. Under the smoothed model, for any $B > 0$, we have*

$$\mathbb{P}_\xi\left[0 < \left|\mu^{mp,\mathbf{r}}(v) - \mu^{mp,\mathbf{r}}(v')\right| < B\right] \le 2Bmn\phi.$$

Proof sketch. With probability 1 the Blackwell optimal pair is unique; fix it and let $C^{\mathbf{r}}_v$ be the recurrent cycle reached from v, of average reward $\mu^{\mathrm{mp},\mathbf{r}}(v)$. If $\mu^{\mathrm{mp},\mathbf{r}}(v) > \mu^{\mathrm{mp},\mathbf{r}}(v')$, then no edge of $C^{\mathbf{r}}_v$ is reachable from v' under the optimal Min strategy $\pi^{\mathrm{bw}}_{\min}$, since otherwise Max could reach and repeat $C^{\mathbf{r}}_v$ from v' and secure $\mu^{\mathrm{mp},\mathbf{r}}(v) > \mu^{\mathrm{mp},\mathbf{r}}(v')$ against $\pi^{\mathrm{bw}}_{\min}$. Assume $0 < \mu^{\mathrm{mp},\mathbf{r}}(v) - \mu^{\mathrm{mp},\mathbf{r}}(v') <$

B, fix an edge $a \in E(C_v^{\mathbf{r}})$, condition on $\mathbf{r}_{-a}$, and write $x \triangleq \mathbf{r}(a)$. Since a is unreachable from v', the value $\mu^{\mathrm{mp},(x;\mathbf{r}_{-a})}(v')$ is a constant α independent of x, whereas $x \mapsto \mu^{\mathrm{mp},(x;\mathbf{r}_{-a})}(v)$ is nondecreasing with derivative at least $1/n$ almost everywhere on the set where a lies on the cycle reached from v. This set therefore maps into $(\alpha, \alpha + B)$ and has Lebesgue measure at most nB; multiplying by the density bound ϕ, summing over the m edges, and adding the symmetric case gives $2mnB\phi$. □

7 Discounted-sum games

In this section, we focus our attention on discounted-sum games under the smoothed model. To establish smoothed polynomial runtime, we first derive a tail bound on the Blackwell threshold $\gamma_{\mathrm{bw}}(\mathbf{r})$ by combining the mathematical ingredients from Section 6 (Section 7.1). We then introduce a restarting variant of policy iteration and establish a total iteration bound for it in terms of $\gamma_{\mathrm{bw}}(\mathbf{r})$ (Section 7.2). Combining this bound with the tail bound yields smoothed polynomial runtime with high probability (Section 7.3). Combining it with a bound on the expected truncated inverse gap yields smoothed polynomial expected runtime (Section 7.4). The full proofs of the results in this section are provided in Appendix D.

7.1 Tail bound

In this subsection, we establish a tail bound on the Blackwell threshold $\gamma_{\mathrm{bw}}(\mathbf{r})$ under the smoothed model. The resulting bound shows that $\gamma_{\mathrm{bw}}(\mathbf{r})$ is bounded away from 1 with high probability and serves as the foundation for both the high probability and expected runtime analyses in the sequel. We first define the variable K which is a polynomial in the input parameters m, n, and ϕ:

$$K \triangleq 92m^2n^2(\phi + 1). \tag{3}$$

Lemma 6 (Tail bound)**.** *Let G be any game graph. Under the smoothed model, for any $x \geq 1$, with the bounding function defined as $M(x) \triangleq 1 + \frac{1}{\theta}\ln(mx)$, we have*

$$\mathbb{P}_\xi\left[\gamma_{bw}(\mathbf{r}) > 1 - \tfrac{1}{x}\right] \leq \frac{KM(x)}{x}.$$

Proof sketch. We combine the reward tail bound Eq. (1) with the suboptimality gap and mean-payoff-value separation events from Lemma 4 and Lemma 5. Condition on the intersection of these events, Lemma 2 and Lemma 3 imply that every non-optimal edge retains its sign for all $\gamma > 1 - 1/x$, and hence $\gamma_{\mathrm{bw}}(\mathbf{r}) \leq 1 - 1/x$. Choosing the separation parameters proportional to $M(x)/x$ and applying union bounds over the edges shows that the complement of this event has probability at most $KM(x)/x$. □

7.2 Restarting Policy Iteration

In this subsection, we introduce the restarting variant of policy iteration called RePI, presented in full detail in Algorithm 2 in Appendix B. We bound its total iteration count in Lemma 7: deterministically by the number of positional strategies for every reward vector, and, under the smoothed model, by an inverse-polynomial factor of the Blackwell gap $1-\gamma_{\mathrm{bw}}(\mathbf{r})$ with probability 1.

Informal description. Our algorithm sweeps through a geometric sequence of thresholds $\gamma_k = 1 - e^{-k}$ for $k = 0, 1, \ldots, \lceil n \ln n \rceil$. For each k in turn, if the target discount factor γ already satisfies $\gamma \leq \gamma_k$, the algorithm directly executes the standard PI procedure on γ. Otherwise, it solves the game with the discount factor γ_k to obtain a candidate strategy pair, computes the payoff of this pair under the target discount factor γ, and checks for any unilateral improving switches. If none exists, the candidate is optimal for the discount factor γ and the algorithm terminates. If the test fails, the algorithm proceeds to the next threshold γ_{k+1} and retries. If every threshold in the sequence fails, then the algorithm falls back to running the standard PI procedure on γ.

Lemma 7 (Total iteration bound)**.** *Let G be any game graph with $n \geq 2$, let $\gamma \in [0, 1)$, and let $\mathbf{r} \in \mathbb{R}^E$. Write $N_{iter}(\mathbf{r})$ for the total number of iterations executed by $\mathsf{RePI}(G, \gamma, \mathbf{r})$ (Algorithm 2), summed over all calls to PI. Then $N_{iter}(\mathbf{r}) \leq (\lceil n \ln n \rceil + 2)\, n^n$ for every $\mathbf{r}$, and, under the smoothed model, with probability 1,*

$$N_{iter}(\mathbf{r}) \leq 27 \cdot \frac{m}{1 - \gamma_{bw}(\mathbf{r})} \ln\Big(\frac{e \cdot n}{1 - \gamma_{bw}(\mathbf{r})}\Big).$$

7.3 High Probability Guarantees

In this subsection, we combine the total iteration bound of Lemma 7 with the tail bound from Section 7.1 to show that, with probability at least $1-\epsilon$, RePI terminates after a number of iterations polynomial in the game parameters, the density bound, the inverse of tail bound, and the inverse failure probability $1/\epsilon$.

Theorem 8. *The restarting PI algorithm achieves smoothed polynomial time with high probability guarantee for discounted-sum games.*

Proof sketch. Fix a failure probability $\epsilon \in (0, 1)$. The tail bound in Lemma 6 shows that the inverse Blackwell gap $1/(1 - \gamma_{\mathrm{bw}}(\mathbf{r}))$ is bounded by a polynomial in $n, m, \phi, 1/\theta$, and $1/\epsilon$, with probability at least $1 - \epsilon$. Conditioned on this event, substituting this bound into the total iteration bound of Lemma 7 yields a polynomial bound on the total number of iterations executed by RePI. The full proof is given in Lemma 16 in Appendix D. □

7.4 Expected Runtime Guarantees

To bound the expected number of iterations of RePI, we first define the notion of truncated inverse gap and show that its expectation is bounded by a polynomial in the game parameters, the density bound, and the inverse of the tail bound in Lemma 9 (Ingredient 4 from Section 4). We then combine this bound with the total iteration bound of Lemma 7 to obtain the desired polynomial bound on the expected number of iterations.

Truncated inverse gap. For any reward vector $\mathbf{r} \in \mathbb{R}^E$, define

$$Y(\mathbf{r}) \triangleq \min\left(1/(1 - \gamma_{\mathrm{bw}}(\mathbf{r})) \ln\left(1/(1 - \gamma_{\mathrm{bw}}(\mathbf{r}))\right),\, n^n\right).$$

Lemma 9 (Bounding expected truncated inverse gap)**.** *Let G be any game graph with $n \geq 2$. Let K be defined in Eq. (3). Under the smoothed model, we have*

$$\mathbb{E}_\xi\big[Y(\mathbf{r})\big] \leq \frac{6(\theta + 1)}{\theta} K (n \ln m)^3.$$

Proof sketch. By Lemma 6, the inverse Blackwell gap has tail probability of $\mathcal{O}(\ln(x)/x)$. Integrating this tail up to the truncation point n^n yields the stated polynomial bound. □

Theorem 10. *The restarting PI algorithm achieves smoothed polynomial time with expected runtime guarantee for discounted-sum games.*

Proof sketch. By Lemma 7, the runtime of RePI is bounded by the inverse Blackwell gap (up to polynomial factors of n and m), while its worst-case fallback is bounded by the finite number of strategies. These two bounds are captured by the truncated inverse gap $Y(\mathbf{r})$. Therefore, the polynomial expectation of $Y(\mathbf{r})$ from Lemma 9 yields a polynomial expected runtime. The full proof is given in Lemma 17 in Appendix D. □

As established in Section 5, any Blackwell optimal strategy for a discounted-sum game is an optimal strategy for the corresponding mean-payoff game [BK76]. Therefore, our smoothed analysis for computing Blackwell optimal strategies directly yields a smoothed polynomial-time algorithm for mean-payoff games, presented in Corollary 11. The full proof is given in Appendix E.

Corollary 11. *The restarting mean-payoff PI algorithm* RePI_{mp} *achieves smoothed polynomial time for mean-payoff games, with both high-probability and expected runtime guarantees.*

Acknowledgement

The authors are thankful for the early discussion with Raimundo Saona and Elahe Tohidi. This paper receives funding from ERC CoG 863818 (ForM-SMArt) and Austrian Science Fund (FWF) 10.55776/COE12.

References

[ACSSU24] Ali Asadi, Krishnendu Chatterjee, Jakub Svoboda, and Raimundo Saona Urmeneta. Deterministic sub-exponential algorithm for discounted-sum games with unary weights. In *Proceedings of the 39th Annual ACM/IEEE Symposium on Logic in Computer Science*, pages 6:1–6:12, 2024. 1

[BCHJ09] Roderick Bloem, Krishnendu Chatterjee, Thomas A Henzinger, and Barbara Jobstmann. Better quality in synthesis through quantitative objectives. In *International Conference on Computer Aided Verification*, pages 140–156. Springer, 2009. 1

[BEF+11] Endre Boros, Khaled Elbassioni, Mahmoud Fouz, Vladimir Gurvich, Kazuhisa Makino, and Bodo Manthey. Stochastic mean payoff games: Smoothed analysis and approximation schemes. In *International Colloquium on Automata, Languages, and Programming*, pages 147–158. Springer, 2011. 1

[BK76] Truman Bewley and Elon Kohlberg. The asymptotic theory of stochastic games. *Mathematics of Operations Research*, 1(3):197–208, 1976. 5, 6.1, 7.4, C.1, C.1, C.3, E.1, E.1

[Bla62] David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33(2):719–726, 1962. 5, C.1

[BV07] Henrik Björklund and Sergei Vorobyov. A combinatorial strongly subexponential strategy improvement algorithm for mean payoff games. *Discrete Applied Mathematics*, 155(2):210–229, 2007. 1, E.1, 9, E.1

[CIJ14] Krishnendu Chatterjee and Rasmus Ibsen-Jensen. The complexity of ergodic mean-payoff games. In *International Colloquium on Automata, Languages, and Programming*, pages 122–133. Springer, 2014. 1

[CKS81] Ashok K Chandra, Dexter C Kozen, and Larry J Stockmeyer. Alternation. *Journal of the ACM (JACM)*, 28(1):114–133, 1981. 1

[CLSZ25] Krishnendu Chatterjee, David Lurie, Raimundo Saona, and Bruno Ziliotto. Uniform value and decidability in ergodic blind stochastic games. *Mathematics of Operations Research*, 2025. in press. 1

[Con92] Anne Condon. The complexity of stochastic games. *Information and Computation*, 96(2):203–224, 1992. 1, 2, 2, 5

[CPV15] Krishnendu Chatterjee, Andreas Pavlogiannis, and Yaron Velner. Quantitative interprocedural analysis. In *Proceedings of the 42nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 539–551, 2015. 1

[CY23] Miranda Christ and Mihalis Yannakakis. The smoothed complexity of policy iteration for markov decision processes. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1890–1903, 2023. 1

[DAHM03] Luca De Alfaro, Thomas A Henzinger, and Rupak Majumdar. Discounting the future in systems theory. In *International Colloquium on Automata, Languages, and Programming*, pages 1022–1037. Springer, 2003. 1

[DKZ19] Dani Dorfman, Haim Kaplan, and Uri Zwick. A faster deterministic exponential time algorithm for energy games and mean payoff games. In *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, pages 114:1–114:14. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2019. 1

[EM79] Andrzej Ehrenfeucht and Jan Mycielski. Positional strategies for mean payoff games. *International Journal of Game Theory*, 8(2):109–113, 1979. 1, 2

[FV96] Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer, 1996. 1, 5

[Gil57] Dean Gillette. Stochastic games with zero stop probabilities. *Contributions to the Theory of Games III*, 39:179–187, 1957. 1

[GKK88] Vladimir A Gurvich, Alexander V Karzanov, and Leonid G Khachiyan. Cyclic games and an algorithm to find minimax cycle means in directed graphs. *USSR Computational Mathematics and Mathematical Physics*, 28(5):85–91, 1988. 1

[HK66] Alan J Hoffman and Richard M Karp. On nonterminating stochastic games. *Management Science*, 12(5):359–370, 1966. 1

[HMZ13] Thomas Dueholm Hansen, Peter Bro Miltersen, and Uri Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16, 2013. 1

[HZ98] Eric A Hansen and Shlomo Zilberstein. Heuristic search in cyclic and/or graphs. In *AAAI/IAAI*, pages 412–418, 1998. 1

[Koz21] Alexander Kozachinskiy. Polyhedral value iteration for discounted games and energy games. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 600–616. SIAM, 2021. 1

[LS24] Bruno Loff and Mateusz Skomra. Smoothed analysis of deterministic discounted and mean-payoff games. In *51st International Colloquium on Automata, Languages, and Programming (ICALP 2024)*, pages 147:1–147:16. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2024. 1, 4, B

[Lud95] Walter Ludwig. A subexponential randomized algorithm for the simple stochastic game problem. *Information and computation*, 117(1):151–155, 1995. 1

[MB85] Ambuj Mahanti and Amitava Bagchi. And/or graph heuristic search methods. *Journal of the ACM (JACM)*, 32(1):28–51, 1985. 1

[MHC03] Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147(1-2):5–34, 2003. 1

[PR89] Amir Pnueli and Roni Rosner. On the synthesis of a reactive module. In *Proceedings of the 16th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 179–190, 1989. 1

[Put94] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 1994. 1

[PY15] Ian Post and Yinyu Ye. The simplex method is strongly polynomial for deterministic markov decision processes. *Mathematics of Operations Research*, 40(4):859–868, 2015. 5, 5

[RW87] Peter J Ramadge and W Murray Wonham. Supervisory control of a class of discrete event processes. *SIAM journal on control and optimization*, 25(1):206–230, 1987. 1

[Sha53] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953. 1

[ST04] Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004. 1, 3

[ST09] Daniel A Spielman and Shang-Hua Teng. Smoothed analysis: an attempt to explain the behavior of algorithms in practice. *Communications of the ACM*, 52(10):76–84, 2009. 1

[ZP96] Uri Zwick and Mike Paterson. The complexity of mean payoff games on graphs. *Theoretical Computer Science*, 158(1-2):343–359, 1996. 1

A Notation summary

Table 1 collects the recurring notation of the paper. Symbols are grouped by role; entries point to the section or equation where they are first introduced.

Game graph and games	
$G = (V_{\min}, V_{\max}, E)$	game graph; $V = V_{\min} \uplus V_{\max}$, $n = \lvert V\rvert$, $m = \lvert E\rvert$
$G^{\gamma,\mathbf{r}}$	discounted-sum game on G with discount γ and rewards $\mathbf{r}$
$G^{\mathrm{mp},\mathbf{r}}$	mean-payoff game on G with rewards $\mathbf{r}$
Strategies	
$\pi_{\min}, \pi_{\max}$	positional strategies of the Min / Max player
$\Pi_{\min}, \Pi_{\max}$	sets of all positional strategies
$\pi^{\mathrm{bw}}_{\min}, \pi^{\mathrm{bw}}_{\max}$	Blackwell optimal strategies (Section 5)
$I_{\min}, I_{\max}$	vertices with improving switches for Min / Max
Rewards, perturbation, values	
$\mathbf{r}_0 \in [-1,1]^E$	adversarial (unperturbed) reward vector
$\xi \in \mathbb{R}^E$	noise vector; density bounded by ϕ, tail parameter θ
$\mathbf{r} = \mathbf{r}_0 + \xi \in \mathbb{R}^E$	perturbed reward vector (Section 3)
$\gamma \in [0,1)$	discount factor
$\mu^{\gamma,\mathbf{r}}(v)$, $\mu^{\mathrm{mp},\mathbf{r}}(v)$	discounted-sum / mean-payoff value at vertex v
$\gamma_{\mathrm{bw}}(\mathbf{r})$	Blackwell threshold (Section 5)
Algorithms	
$\mathsf{PI}(G,\gamma,\mathbf{r})$	policy iteration on $G^{\gamma,\mathbf{r}}$ (Section 5)
RePI	restarting policy iteration for discounted-sum games (Algorithm 2)
$\mathsf{RePI}_{\mathrm{mp}}$	restarting policy iteration for mean-payoff games (Algorithm 3)
$\gamma_k = 1 - e^{-k}$	geometric threshold schedule used by RePI / $\mathsf{RePI}_{\mathrm{mp}}$
Analysis quantities	
$\Delta^{\gamma,\mathbf{r}}(v,v')$	suboptimality gap of the edge (v,v')
$K \triangleq 92m^2n^2(\phi+1)$	coefficient in the tail bound (Eq. (3))
$M(x) \triangleq 1 + \frac{1}{\theta}\ln(mx)$	tail bound function (Lemma 6)
$Y(\mathbf{r})$	truncated inverse gap (Section 7.4)
$\epsilon \in (0,1)$	failure probability for high-probability guarantees

Table 1: Recurring notation used throughout the paper.

B Algorithm Pseudocodes

In this section, we present the pseudocodes of the policy iteration algorithms for discounted-sum games discussed in the main body: the classic policy iteration PI (Algorithm 1) and the restarting variant RePI (Algorithm 2). The second algorithm, Algorithm 2, is related in spirit to the increasing-discount algorithm IncreasingDiscountPI2 of [LS24], but the crucial difference is that Algorithm 2 has an explicit fallback after a fixed finite sequence of thresholds. This gives a uniform deterministic bound on the running time, which is essential for our expected-runtime analysis. Thus, although IncreasingDiscountPI2 yields a high-probability smoothed guarantee in [LS24], the

algorithm does not achieve a polynomial expected-running-time guarantee based on our techniques. Moreover, IncreasingDiscountPI2 warm-starts policy iteration from the policies computed at the previous discount factor, whereas Algorithm 2 restarts policy iteration from scratch at each threshold; this difference, however, is not used in our analysis.

Algorithm 1 Policy Iteration: $\mathsf{PI}(G, \gamma, \mathbf{r})$

Input: Graph $G = (V_{\min}, V_{\max}, E)$, discount factor $\gamma \in [0, 1)$, reward vector $\mathbf{r} \in \mathbb{R}^E$.
Output: Optimal strategy pair $(\pi^*_{\min}, \pi^*_{\max})$ and value vector $\mu^{\gamma,\mathbf{r}}$ for the game $G^{\gamma,\mathbf{r}}$.
Initialize an arbitrary positional strategy $\pi_{\min} \in \Pi_{\min}$ for the Min Player.
loop
    // Policy Evaluation Phase
    Compute the value vector $\mu \in \mathbb{R}^V$ and Max Player's best-response $\pi_{\max} \in \Pi_{\max}$ satisfying:
    $\mu(v) = \mathbf{r}(v, \pi_{\min}(v)) + \gamma\mu(\pi_{\min}(v))$ for all $v \in V_{\min}$,
    $\mu(v) = \max_{(v,v') \in E}\big(\mathbf{r}(v, v') + \gamma\mu(v')\big)$ for all $v \in V_{\max}$.
    // Policy Improvement Phase
    Let $I = \big\{v \in V_{\min} \mid \exists (v, v') \in E$ such that $\mathbf{r}(v, v') + \gamma\mu(v') < \mu(v)\big\}$.
    if $I = \varnothing$ **then**
        return $(\pi_{\min}, \pi_{\max}, \mu)$
    end if
    Update $\pi_{\min}$ by switching the edge for **all** vertices $v \in I$ to minimize the value:
    $\pi_{\min}(v) \leftarrow \arg\min_{(v,v') \in E}\big(\mathbf{r}(v, v') + \gamma\mu(v')\big)$.
end loop

Algorithm 2 Restarting Policy Iteration: $\mathsf{RePI}(G, \gamma, \mathbf{r})$

Input: Graph $G = (V_{\min}, V_{\max}, E)$ with $n = |V| \geq 2$, target discount factor $\gamma \in [0, 1)$, reward vector $\mathbf{r} \in \mathbb{R}^E$.
Output: Optimal strategy pair $(\pi^*_{\min}, \pi^*_{\max})$ and value vector $\mu^{\gamma,\mathbf{r}}$ for the game $G^{\gamma,\mathbf{r}}$.
Let $k_{\max} = \lceil n \ln n \rceil$.
for $k = 0, 1, \ldots, k_{\max}$ **do**
    Set the threshold discount factor $\gamma_k = 1 - e^{-k}$.
    if $\gamma \leq \gamma_k$ **then**
        return $\mathsf{PI}(G, \gamma, \mathbf{r})$
    else
        // Phase 1: Solve at the threshold
        $(\pi_{\min}, \pi_{\max}, \mu_{\gamma_k}) \leftarrow \mathsf{PI}(G, \gamma_k, \mathbf{r})$
        // Phase 2: Optimality test for the target discount factor
        Compute the value vector $\mu_\gamma \in \mathbb{R}^V$ by evaluating $(\pi_{\min}, \pi_{\max})$ under γ.
        Let $I_{\min} = \big\{v \in V_{\min} \mid \exists (v, v') \in E$ such that $\mathbf{r}(v, v') + \gamma\mu_\gamma(v') < \mu_\gamma(v)\big\}$.
        Let $I_{\max} = \big\{v \in V_{\max} \mid \exists (v, v') \in E$ such that $\mathbf{r}(v, v') + \gamma\mu_\gamma(v') > \mu_\gamma(v)\big\}$.
        if $I_{\min} = \varnothing$ **and** $I_{\max} = \varnothing$ **then**
            return $(\pi_{\min}, \pi_{\max}, \mu_\gamma)$
        end if
    end if
end for
// Phase 3: Fallback
return $\mathsf{PI}(G, \gamma, \mathbf{r})$

C Proofs of Section 6

In this section, we provide the proofs omitted from Section 6, organized by the same ingredient structure as the main body.

C.1 Ingredient 1: Value Decomposition and Lipschitz Continuity

Lemma 2 (Value decomposition under Blackwell optimality)**.** *Consider a game graph G with reward vector $\mathbf{r} \in \mathbb{R}^E$. Let $v \in V$ be a vertex.*

1. Decomposition: *For all $\gamma \in (\gamma_{bw}(\mathbf{r}), 1)$, the discounted-sum value admits the following decomposition, where $w_v^{\mathbf{r}}(\gamma)$ is a rational function of γ:*
$$\mu^{\gamma,\mathbf{r}}(v) = \frac{\mu^{mp,\mathbf{r}}(v)}{1-\gamma} + w_v^{\mathbf{r}}(\gamma).$$
2. Magnitude and derivative: *For all $\gamma \in (\gamma_{bw}(\mathbf{r}), 1)$, the magnitude and the derivative of $w_v^{\mathbf{r}}(\gamma)$ are bounded by:*
$$|w_v^{\mathbf{r}}(\gamma)| \le 4n\|\mathbf{r}\|_\infty \quad \textit{and} \quad \left|(\frac{d}{d\gamma} w_v^{\mathbf{r}})(\gamma)\right| \le 7n^2\|\mathbf{r}\|_\infty.$$

Proof. Let $M = \|\mathbf{r}\|_\infty$ and $\mu \triangleq \mu^{\mathrm{mp},\mathbf{r}}(v)$. Since $\gamma \in (\gamma_{\mathrm{bw}}(\mathbf{r}), 1)$, we consider a Blackwell optimal pair of positional strategies $(\pi^{\mathrm{bw}}_{\min}, \pi^{\mathrm{bw}}_{\max})$ for $G^{\gamma,\mathbf{r}}$. Observe that the path from vertex v under the strategy profile $(\pi^{\mathrm{bw}}_{\min}, \pi^{\mathrm{bw}}_{\max})$ consists of a finite path of length $t \le n$ and a recurrent cycle of length $c \le n$. Let r_k denote the reward at step k. Note that the mean-payoff is the cycle's average reward, i.e., $\mu = \frac{1}{c}\sum_{j=0}^{c-1} r_{t+j}$. Therefore, we have $|\mu| \le M$ and $\sum_{j=0}^{c-1}(r_{t+j} - \mu) = 0$. We now present the two items of the proof.

Value decomposition. The value vector expands as follows:

$$\begin{aligned}
\mu^{\gamma,\mathbf{r}}(v) &= \sum_{j=0}^{t-1} r_j\gamma^j + \frac{\gamma^t}{1-\gamma^c}\sum_{j=0}^{c-1} r_{t+j}\gamma^j \\
&= \sum_{j=0}^{t-1}(r_j-\mu)\gamma^j + \frac{\gamma^t}{1-\gamma^c}\sum_{j=0}^{c-1}(r_{t+j}-\mu)\gamma^j + \mu\sum_{j=0}^{t-1}\gamma^j + \frac{\gamma^t\mu}{1-\gamma^c}\sum_{j=0}^{c-1}\gamma^j \\
&= \sum_{j=0}^{t-1}(r_j-\mu)\gamma^j + \frac{\gamma^t}{1-\gamma^c}\sum_{j=0}^{c-1}(r_{t+j}-\mu)(\gamma^j-1) + \mu\Big(\frac{1-\gamma^t}{1-\gamma} + \frac{\gamma^t}{1-\gamma}\Big) \\
&= \underbrace{\sum_{j=0}^{t-1}(r_j-\mu)\gamma^j + \sum_{j=1}^{c-1}\Big(\frac{\sum_{i=0}^{j-1}\gamma^i}{\sum_{i=0}^{c-1}\gamma^i}\Big)(\mu - r_{t+j})\gamma^t}_{\triangleq w_v^{\mathbf{r}}(\gamma)} + \frac{\mu}{1-\gamma},
\end{aligned}$$

where the first equality is due to decomposing the path into a transient path and a cycle, the second equality rearranges the terms to separate the mean-payoff component, the third equality introduces -1 into the cycle sum because $\sum_{j=0}^{c-1}(r_{t+j}-\mu) = 0$, and the final equality factors out $1-\gamma$. We define $f_j(\gamma) \triangleq \sum_{i=0}^{j-1}\gamma^i$. Then, we rewrite $w_v^{\mathbf{r}}(\gamma)$ as:

$$\begin{aligned}
w_v^{\mathbf{r}}(\gamma) &= \sum_{j=0}^{t-1}(r_j-\mu)\gamma^j + \frac{1}{f_c(\gamma)}\sum_{j=1}^{c-1}(\mu - r_{t+j})\gamma^t f_j(\gamma) \\
&= \frac{\sum_{j=0}^{t-1}(r_j-\mu)\gamma^j f_c(\gamma) + \sum_{j=1}^{c-1}(\mu - r_{t+j})\gamma^t f_j(\gamma)}{f_c(\gamma)}.
\end{aligned}$$

Therefore, $w_v^{\mathbf{r}}(\gamma)$ is a rational function of γ, which proves the first item of the statement. We now bound its magnitude and derivative.

Magnitude bound. For magnitude, we get

$$\begin{aligned}
|w_v^{\mathbf{r}}(\gamma)| &= \Big|\sum_{j=0}^{t-1}(r_j-\mu)\gamma^j + \sum_{j=1}^{c-1}(\mu - r_{t+j})\gamma^t \frac{f_j(\gamma)}{f_c(\gamma)}\Big| \\
&\le \sum_{j=0}^{t-1}|r_j-\mu|\gamma^j + \sum_{j=1}^{c-1}|\mu - r_{t+j}|\gamma^t \frac{f_j(\gamma)}{f_c(\gamma)} \\
&\le 2M\sum_{j=0}^{t-1}\gamma^j + 2M\sum_{j=1}^{c-1}\gamma^t\frac{f_j(\gamma)}{f_c(\gamma)} \\
&\le 2M\sum_{j=0}^{t-1}1 + 2M\sum_{j=1}^{c-1}\frac{f_j(\gamma)}{f_c(\gamma)} \\
&\le 4Mn = 4n\|\mathbf{r}\|_\infty,
\end{aligned}$$

where the first inequality is due to the triangle inequality, the second inequality is due to the bounds $|r_j-\mu| \le 2M$ and $|\mu - r_{t+j}| \le 2M$, the third inequality is due to $\gamma^j, \gamma^t \le 1$ for all $\gamma \in [0,1]$, the fourth inequality is due to $f_j(\gamma) \le f_c(\gamma)$ for all $j < c$, and the final inequality is due to $t, c \le n$.

Derivative bound. For the derivative, we first compute $\Big|\frac{d}{d\gamma}\frac{f_j(\gamma)}{f_c(\gamma)}\Big|$:

$$\begin{aligned}
\Big|\frac{d}{d\gamma}\frac{f_j(\gamma)}{f_c(\gamma)}\Big| &= \Big|\frac{(\frac{d}{d\gamma}f_j)(\gamma)f_c(\gamma) - f_j(\gamma)(\frac{d}{d\gamma}f_c)(\gamma)}{f_c(\gamma)^2}\Big| \\
&= \Big|\frac{(\frac{d}{d\gamma}f_j)(\gamma)}{f_c(\gamma)} - \frac{f_j(\gamma)}{f_c(\gamma)}\cdot\frac{(\frac{d}{d\gamma}f_c)(\gamma)}{f_c(\gamma)}\Big| \\
&\le \Big|\frac{(\frac{d}{d\gamma}f_j)(\gamma)}{f_c(\gamma)}\Big| + \Big|\frac{(\frac{d}{d\gamma}f_c)(\gamma)}{f_c(\gamma)}\Big| \\
&\le 2n,
\end{aligned}$$

where the first equality is due to the quotient rule, the second equality is due to rearranging the terms, the first inequality is due to the triangle inequality and the fact that $|\frac{f_j(\gamma)}{f_c(\gamma)}| \le 1$, and the second inequality is due to $(\frac{d}{d\gamma}f_j)(\gamma) = \sum_{i=1}^{j-1} i\gamma^{i-1} \le \sum_{i=1}^{c-1} i\gamma^{i-1} = (\frac{d}{d\gamma}f_c)(\gamma) \le c\sum_{i=1}^{c-1}\gamma^{i-1} \le nf_c(\gamma)$. Now, we bound the derivative of $w_v^{\mathbf{r}}(\gamma)$ as follows:

$$\begin{aligned}
\Big|\frac{d}{d\gamma}w_v^{\mathbf{r}}(\gamma)\Big| &\le \sum_{j=0}^{t-1}|r_j-\mu|\Big|\frac{d}{d\gamma}\gamma^j\Big| + \sum_{j=1}^{c-1}|\mu-r_{t+j}|\Big|\frac{d}{d\gamma}\Big(\gamma^t\frac{f_j(\gamma)}{f_c(\gamma)}\Big)\Big| \\
&\le 2M\sum_{j=0}^{t-1}\Big|\frac{d}{d\gamma}\gamma^j\Big| + 2M\sum_{j=1}^{c-1}\Big|\frac{d}{d\gamma}\Big(\gamma^t\frac{f_j(\gamma)}{f_c(\gamma)}\Big)\Big| \\
&\le 2M\sum_{j=1}^{t-1}j + 2M\sum_{j=1}^{c-1}\Big|\frac{d}{d\gamma}\Big(\gamma^t\frac{f_j(\gamma)}{f_c(\gamma)}\Big)\Big| \\
&\le Mn^2 + 2M\sum_{j=1}^{c-1}3n \\
&\le 7n^2M = 7n^2\|\mathbf{r}\|_\infty,
\end{aligned}$$

where the first inequality is due to the triangle inequality, the second inequality is due to $|r_j - \mu| \leq 2M$ and $|\mu - r_{t+j}| \leq 2M$, the third inequality is due to $\left|\frac{d}{d\gamma}\gamma^j\right| = j\gamma^{j-1} \leq j$ for all $\gamma \in [0,1]$, the fourth inequality is due to $t \leq n$ and $\left|\frac{d}{d\gamma}\frac{f_j(\gamma)}{f_c(\gamma)}\right| \leq 2n$ as shown above, and the final inequality is due to $c \leq n$.

Hence all the desired items are proved and the result follows. □

Lemma 3 (Lipschitz continuity)**.** *Let $\mathbf{r} \in \mathbb{R}^E$ be any reward vector. For any edge $a = (v, v') \in E$ where $\mu^{mp,\mathbf{r}}(v) = \mu^{mp,\mathbf{r}}(v')$, the map $\gamma \mapsto \Delta^{\gamma,\mathbf{r}}(v, v')$ is $18n^2\|\mathbf{r}\|_\infty$-Lipschitz continuous over the interval $(\gamma_{bw}(\mathbf{r}), 1)$.*

Proof. By Lemma 2, for all $\gamma \in [\gamma_{\text{bw}}(\mathbf{r}), 1)$, the value at vertices $v, v' \in V$ decomposes as:

$$\mu^{\gamma,\mathbf{r}}(v) = \frac{\mu^{\text{mp},\mathbf{r}}(v)}{1-\gamma} + w_v^{\mathbf{r}}(\gamma)\,, \quad \text{and} \quad \mu^{\gamma,\mathbf{r}}(v') = \frac{\mu^{\text{mp},\mathbf{r}}(v')}{1-\gamma} + w_{v'}^{\mathbf{r}}(\gamma)\,,$$

where $w_v^{\mathbf{r}}(\gamma)$ and $w_{v'}^{\mathbf{r}}(\gamma)$ are rational functions of γ with bounded magnitude and derivative as stated in Lemma 2.

Substituting this decomposition into the suboptimality gap yields:

$$\begin{aligned}\Delta^{\gamma,\mathbf{r}}(v, v') &= \mathbf{r}(v, v') + \gamma\Big(\frac{\mu^{\text{mp},\mathbf{r}}(v')}{1-\gamma} + w_{v'}^{\mathbf{r}}(\gamma)\Big) - \Big(\frac{\mu^{\text{mp},\mathbf{r}}(v)}{1-\gamma} + w_v^{\mathbf{r}}(\gamma)\Big) \\ &= \mathbf{r}(v, v') - \mu^{\text{mp},\mathbf{r}}(v) + \gamma w_{v'}^{\mathbf{r}}(\gamma) - w_v^{\mathbf{r}}(\gamma)\,,\end{aligned}$$

where the first equality is due to substituting the value decomposition, and the second equality is due to the assumption that $\mu^{\text{mp},\mathbf{r}}(v) = \mu^{\text{mp},\mathbf{r}}(v')$. Notice that the singular term $\frac{\mu^{\text{mp},\mathbf{r}}(v)}{1-\gamma}$ exactly cancels out. To bound the Lipschitz constant over $[\gamma_{\text{bw}}(\mathbf{r}), 1)$, we evaluate the derivative with respect to γ:

$$\frac{d}{d\gamma}\Delta^{\gamma,\mathbf{r}}(v, v') = w_{v'}^{\mathbf{r}}(\gamma) + \gamma(\frac{d}{d\gamma}w_{v'}^{\mathbf{r}})(\gamma) - (\frac{d}{d\gamma}w_v^{\mathbf{r}})(\gamma).$$

Therefore, we obtain:

$$\begin{aligned}\Big|\frac{d}{d\gamma}\Delta^{\gamma,\mathbf{r}}(v, v')\Big| &\leq |w_{v'}^{\mathbf{r}}(\gamma)| + |(\frac{d}{d\gamma}w_{v'}^{\mathbf{r}})(\gamma)| + |(\frac{d}{d\gamma}w_v^{\mathbf{r}})(\gamma)| \\ &\leq 4n\|\mathbf{r}\|_\infty + 7n^2\|\mathbf{r}\|_\infty + 7n^2\|\mathbf{r}\|_\infty \leq 18n^2\|\mathbf{r}\|_\infty\,,\end{aligned}$$

where the first inequality is due to the triangle inequality and the fact that $\gamma < 1$, and the second inequality is due to the bounds on the magnitude and derivative of $w_v^{\mathbf{r}}(\gamma)$ and $w_{v'}^{\mathbf{r}}(\gamma)$ as stated in Lemma 2. Hence, the map $\gamma \mapsto \Delta^{\gamma,\mathbf{r}}(v, v')$ is Lipschitz continuous over $[\gamma_{\text{bw}}(\mathbf{r}), 1)$ with Lipschitz constant $18n^2\|\mathbf{r}\|_\infty$. □

We conclude by showing that under the smoothed model, with probability one, every vertex has a unique outgoing edge with zero suboptimality gap (Proposition 12), which yields the uniqueness of the Blackwell optimal strategy pair (Corollary 13). These results use standard ideas from smoothed analysis. The following arguments exploit the bounded density of the noise by conditioning on all but one coordinate at a time, which calls for notation that isolates a single edge.

Single-edge decomposition. To analyze the effect of perturbing a single edge, we decompose the reward and noise vectors along any edge $a \in E$. We write

$$\mathbf{r} = (\mathbf{r}(a); \mathbf{r}_{-a}) \quad \text{and} \quad \xi = (\xi_a; \xi_{-a}),$$

where $\mathbf{r}_{-a} \in \mathbb{R}^{E\setminus\{a\}}$ is the vector of rewards for all edges except a, defined by $\mathbf{r}_{-a}(v, v') = \mathbf{r}(v, v')$ for all $(v, v') \in E \setminus \{a\}$, and ξ_{-a} is defined analogously. For any fixed $\mathbf{r}_{-a} \in \mathbb{R}^{E\setminus\{a\}}$, we write $\mathbb{P}_{\xi_a}[\cdot]$ for the probability measure over the single noise coordinate ξ_a.

Proposition 12 (Optimality uniqueness)**.** *Under the smoothed model, we have*

$$\mathbb{P}_\xi\big[\forall v \in V,\ \exists!\, a = (v, v') \in E \text{ such that } \mu^{mp,\mathbf{r}}(v) = \mu^{mp,\mathbf{r}}(v') \text{ and } \Delta^{1,\mathbf{r}}(a) = 0\big] = 1.$$

In words, with probability 1*, for each vertex* $v \in V$*, there exists a unique outgoing edge* $a = (v, v')$ *such that* $\mu^{mp,\mathbf{r}}(v) = \mu^{mp,\mathbf{r}}(v')$ *and* $\Delta^{1,\mathbf{r}}(a) = 0$*.*

Proof. We separately establish (i) existence, i.e., for every vertex $v \in V$ at least one such edge exists deterministically; and (ii) uniqueness, i.e., with probability 1 no two distinct outgoing edges from v both satisfy the conditions. A union bound over the finite vertex set V completes the proof.

Existence. Let $\mathbf{r} \in \mathbb{R}^E$ be any reward vector and $(\pi^{\text{bw}}_{\min}, \pi^{\text{bw}}_{\max})$ be a Blackwell optimal strategy pair. Fix a vertex $v \in V$ and let $a = (v, v') \in E$ be the edge selected at v by this strategy pair. Note that the path from v under the strategy profile $(\pi^{\text{bw}}_{\min}, \pi^{\text{bw}}_{\max})$ consists of a finite path of length $t \le n$ and a recurrent cycle of length $c \le n$. The analogous path from v' contains the same cycle. Since the mean-payoff is the average reward of the cycle and Blackwell optimal strategies are also mean-payoff optimal [BK76], we have $\mu^{\text{mp},\mathbf{r}}(v) = \mu^{\text{mp},\mathbf{r}}(v')$. Moreover, since a is an optimal action, we have $\Delta^{\gamma,\mathbf{r}}(a) = 0$ for all $\gamma \in (\gamma_{\text{bw}}(\mathbf{r}), 1)$. By Lemma 3, the limit extends to $\Delta^{1,\mathbf{r}}(a) = 0$, which yields the existence.

Uniqueness. Fix a vertex $v \in V$ with at least two outgoing edges, otherwise uniqueness holds. Let $a_1 = (v, v_1)$ and $a_2 = (v, v_2)$ be any two distinct outgoing edges from v. We define the event $\mathcal{E}$ that both a_1 and a_2 satisfy the conditions:

$$\mathcal{E} \triangleq \big\{\mathbf{r} \in \mathbb{R}^E : \mu^{\text{mp},\mathbf{r}}(v) = \mu^{\text{mp},\mathbf{r}}(v_1) = \mu^{\text{mp},\mathbf{r}}(v_2),\ \Delta^{1,\mathbf{r}}(a_1) = 0,\ \Delta^{1,\mathbf{r}}(a_2) = 0\big\},$$

where the mean-payoff equality at v_2 is included so that $\Delta^{1,\mathbf{r}}(a_2)$ is well-defined. We show that $\mathbb{P}_\xi[\mathcal{E}] = 0$, which implies that with probability 1 at most one outgoing edge from v satisfies the conditions. Since $a_1 \neq a_2$, a Blackwell optimal strategy must differ from at least one of a_1 or a_2 at vertex v. Therefore, we decompose $\mathcal{E} \subseteq \mathcal{E}_1 \cup \mathcal{E}_2$, where

$$\mathcal{E}_1 \triangleq \mathcal{E} \cap \big\{\mathbf{r} \in \mathbb{R}^E : \exists\, (\pi^{\text{bw}}_{\min}, \pi^{\text{bw}}_{\max}) \text{ Blackwell optimal for } \mathbf{r} \text{ with } \pi^{\text{bw}}(v) \neq v_1\big\},$$
$$\mathcal{E}_2 \triangleq \mathcal{E} \cap \big\{\mathbf{r} \in \mathbb{R}^E : \exists\, (\pi^{\text{bw}}_{\min}, \pi^{\text{bw}}_{\max}) \text{ Blackwell optimal for } \mathbf{r} \text{ with } \pi^{\text{bw}}(v) \neq v_2\big\},$$

and $\pi^{\text{bw}}(v) \triangleq \pi^{\text{bw}}_{\min}(v)$ if $v \in V_{\min}$ and $\pi^{\text{bw}}(v) \triangleq \pi^{\text{bw}}_{\max}(v)$ if $v \in V_{\max}$. We show that $\mathbb{P}_\xi[\mathcal{E}_1] = 0$. The bound $\mathbb{P}_\xi[\mathcal{E}_2] = 0$ follows by the symmetric argument with the roles of a_1 and a_2 swapped. Consider any fixed $\mathbf{r}_{-a_1} \in \mathbb{R}^{E\setminus\{a_1\}}$. We analyze the event $\mathcal{E}_1$ with respect to the random variable $x \triangleq \mathbf{r}(a_1)$. Formally, we prove $\mathbb{P}_{\xi_{a_1}}[\mathcal{E}_1] = 0$. Let $\mu^{\gamma,\mathbf{r}_{-a_1}}$ denote the optimal value vector of the restricted game played on the edge set $E \setminus \{a_1\}$. Since this restricted game is completely independent of a_1, its value vector $\mu^{\gamma,\mathbf{r}_{-a_1}}$ is a function depending only on $\mathbf{r}_{-a_1}$. Indeed, for any $x \in \mathbb{R}$ such that $\mathbf{r} \triangleq (x, \mathbf{r}_{-a_1}) \in \mathcal{E}_1$, there exists a Blackwell optimal strategy pair $(\pi^{\text{bw}}_{\min}, \pi^{\text{bw}}_{\max})$ where $\pi^{\text{bw}}(v) \neq v_1$. Since the edge $(v, \pi^{\text{bw}}(v)) \in E \setminus \{a_1\}$, removing a_1 does not change the optimal value at vertex v. Consequently, the value vectors of the full game and the restricted game are equal, i.e., $\mu^{\gamma,\mathbf{r}} = \mu^{\gamma,\mathbf{r}_{-a_1}}$ for all $\gamma \in (\gamma_{\text{bw}}(\mathbf{r}), 1)$. This equality also implies $\mu^{\text{mp},\mathbf{r}_{-a_1}}(v) = \mu^{\text{mp},\mathbf{r}_{-a_1}}(v_1)$ by Eq. (2). Under the event $\mathcal{E}_1$, taking the limit of suboptimality gap evaluates to:

$$\Delta^{1,\mathbf{r}}(a_1) = x + \lim_{\gamma\uparrow 1} \big(\gamma\mu^{\gamma,\mathbf{r}_{-a_1}}(v_1) - \mu^{\gamma,\mathbf{r}_{-a_1}}(v)\big).$$

Let $C(\mathbf{r}_{-a_1})$ denote the limit term. By Lemma 3 applied to the restricted game, $C(\mathbf{r}_{-a_1})$ exists and is finite. Since $C(\mathbf{r}_{-a_1})$ is fixed with respect to x, the condition $\Delta^{1,\mathbf{r}}(a_1) = 0$ requires $x + C(\mathbf{r}_{-a_1}) = 0$, or equivalently, $x = -C(\mathbf{r}_{-a_1})$. Let $\lambda(\mathcal{E}_1)$ denote the Lebesgue measure of the set of x values satisfying this condition. Note that $\lambda(\mathcal{E}_1) = 0$. Since the probability density of x is uniformly bounded by ϕ, we have $\mathbb{P}_{\xi_{a_1}}[\mathcal{E}_1] = 0$. Taking the expectation over the independent randomness in

ξ_{-a_1} gives

$$\mathbb{P}_\xi[\mathcal{E}_1] = \mathop{\mathbb{E}}_{\xi_{-a_1}}\left[\mathbb{P}_{\xi_{a_1}}[\mathcal{E}_1]\right] = 0.$$

The symmetric argument applied to a_2 gives $\mathbb{P}_\xi[\mathcal{E}_2] = 0$, so a union bound yields $\mathbb{P}_\xi[\mathcal{E}] \le \mathbb{P}_\xi[\mathcal{E}_1] + \mathbb{P}_\xi[\mathcal{E}_2] = 0$. A further union bound over the finitely many pairs of distinct outgoing edges at every vertex $v \in V$ completes the proof. □

Corollary 13 (Uniqueness of Blackwell optimal strategies)**.** *Under the smoothed model, we have*

$$\mathbb{P}_\xi\left[\exists!\,(\pi^{bw}_{\min}, \pi^{bw}_{\max}) \in \Pi_{\min} \times \Pi_{\max} \text{ that is Blackwell optimal for } \mathbf{r}\right] = 1.$$

In words, with probability 1*, there exists a unique Blackwell optimal strategy pair.*

Proof. Existence is guaranteed deterministically by the classical Blackwell optimality result [Bla62, BK76]. It remains to establish uniqueness with probability 1. By Proposition 12, the event

$$\mathcal{E} \triangleq \left\{\mathbf{r} \in \mathbb{R}^E : \forall v \in V,\ \exists!\, a^*(v) = (v, v') \in E \text{ s.t. } \mu^{\mathrm{mp},\mathbf{r}}(v) = \mu^{\mathrm{mp},\mathbf{r}}(v') \text{ and } \Delta^{1,\mathbf{r}}(a^*(v)) = 0\right\}$$

satisfies $\mathbb{P}_\xi[\mathcal{E}] = 1$. We show that on $\mathcal{E}$, the Blackwell optimal strategy pair is unique.

Fix $\mathbf{r} \in \mathcal{E}$ and let $(\pi^{\mathrm{bw}}_{\min}, \pi^{\mathrm{bw}}_{\max})$ be any Blackwell optimal strategy pair for $\mathbf{r}$. For any vertex $v \in V$, let $a = (v, \pi^{\mathrm{bw}}(v)) \in E$ be the edge selected at v by this pair, where $\pi^{\mathrm{bw}}(v) \triangleq \pi^{\mathrm{bw}}_{\min}(v)$ if $v \in V_{\min}$ and $\pi^{\mathrm{bw}}(v) \triangleq \pi^{\mathrm{bw}}_{\max}(v)$ if $v \in V_{\max}$. Since Blackwell optimal strategies are also mean-payoff optimal [BK76], we have $\mu^{\mathrm{mp},\mathbf{r}}(v) = \mu^{\mathrm{mp},\mathbf{r}}(\pi^{\mathrm{bw}}(v))$. Moreover, since Blackwell optimal strategies are optimal in discounted-sum games for all $\gamma \in (\gamma_{\mathrm{bw}}(\mathbf{r}), 1)$, we have $\Delta^{\gamma,\mathbf{r}}(a) = 0$ for such γ, and by Lemma 3, taking the limit yields $\Delta^{1,\mathbf{r}}(a) = 0$. Since $\mathbf{r} \in \mathcal{E}$, the unique edge satisfying these conditions at v is $a^*(v)$, so $a = a^*(v)$. Since this holds for every $v \in V$, both $\pi^{\mathrm{bw}}_{\min}$ and $\pi^{\mathrm{bw}}_{\max}$ are uniquely determined, which completes the proof. □

Lemma 14 (Unique optimal strategies at fixed thresholds)**.** *Let $\Gamma \subseteq [0, 1)$ be any finite set of discount factors. Under the smoothed model, with probability* 1*, for every $\gamma \in \Gamma$ with $\gamma > \gamma_{bw}(\mathbf{r})$, the discounted-sum game $G^{\gamma,\mathbf{r}}$ has a unique optimal strategy pair, and this pair is the unique Blackwell optimal strategy pair for $\mathbf{r}$.*

Proof. Fix a discount factor $\gamma \in [0, 1)$. We first show that, with probability 1, the game $G^{\gamma,\mathbf{r}}$ has a unique optimal strategy pair. It is enough to show that, with probability 1, no vertex has two distinct outgoing edges that both satisfy the discounted Bellman equality $\Delta^{\gamma,\mathbf{r}}(a) = 0$. Indeed, by the Bellman optimality equations, at every vertex at least one outgoing edge satisfies this equality, and if exactly one such edge exists at every vertex, then both players' optimal positional strategies are uniquely determined.

Fix a vertex v with two distinct outgoing edges $a_1 = (v, v_1)$ and $a_2 = (v, v_2)$, and consider the event

$$\mathcal{E} \triangleq \left\{\mathbf{r} \in \mathbb{R}^E : \Delta^{\gamma,\mathbf{r}}(a_1) = 0 \text{ and } \Delta^{\gamma,\mathbf{r}}(a_2) = 0\right\}.$$

We prove that $\mathbb{P}_\xi[\mathcal{E}] = 0$. Condition on all rewards except a_1, writing $x = \mathbf{r}(a_1)$ and $\mathbf{r} = (x; \mathbf{r}_{-a_1})$. On the event $\mathcal{E}$, the edge a_2 is also optimal at v; hence there exists an optimal strategy pair for $G^{\gamma,\mathbf{r}}$ whose strategy at v selects a_2 and therefore does not use a_1. Removing a_1 from the game does not change the value vector at discount γ: if $v \in V_{\min}$, Min still has an optimal strategy that avoids a_1, while removing a Min edge can only increase the value; if $v \in V_{\max}$, Max still has an optimal strategy that avoids a_1, while removing a Max edge can only decrease the value. Thus $\mu^{\gamma,\mathbf{r}} = \mu^{\gamma,\mathbf{r}_{-a_1}}$ on $\mathcal{E}$, where $\mu^{\gamma,\mathbf{r}_{-a_1}}$ denotes the value vector of the game with edge a_1 removed. Consequently, on $\mathcal{E}$,

$$0 = \Delta^{\gamma,\mathbf{r}}(a_1) = x + \gamma\mu^{\gamma,\mathbf{r}_{-a_1}}(v_1) - \mu^{\gamma,\mathbf{r}_{-a_1}}(v).$$

For the fixed conditioned vector $\mathbf{r}_{-a_1}$, the last two terms are independent of x, so this equality can hold for at most one value of x. Since the conditional density of $x = \mathbf{r}(a_1)$ is bounded, the conditional probability of $\mathcal{E}$ is zero. Integrating over $\mathbf{r}_{-a_1}$ gives $\mathbb{P}_\xi[\mathcal{E}] = 0$. A union bound over all vertices and all pairs of distinct outgoing edges shows that $G^{\gamma,\mathbf{r}}$ has a unique optimal strategy pair with probability 1.

Since Γ is finite, another union bound implies that, with probability 1, every game $G^{\gamma,\mathbf{r}}$ with $\gamma \in \Gamma$ has a unique optimal strategy pair. Intersect this event with the probability-one event of Corollary 13. On the intersection, if $\gamma \in \Gamma$ and $\gamma > \gamma_{\mathrm{bw}}(\mathbf{r})$, then the unique Blackwell optimal pair is optimal for $G^{\gamma,\mathbf{r}}$ by the definition of the Blackwell threshold. Since the discounted optimal pair at γ is unique, it must coincide with the unique Blackwell optimal pair. □

C.2 Ingredient 2: Suboptimality Gap Separation

Lemma 4 (Suboptimality separation)**.** *Let G be any game graph and let $a = (v, v') \in E$ be any fixed edge. Under the smoothed model, for any $\delta > 0$, we have*

$$\mathbb{P}_\xi\left[\mu^{mp,\mathbf{r}}(v) = \mu^{mp,\mathbf{r}}(v') \text{ and } 0 < \left|\Delta^{1,\mathbf{r}}(a)\right| < \delta\right] \le 2\delta\phi.$$

Proof. If a is the only outgoing edge from v, it must be selected by any valid strategy. Thus, $\Delta^{1,\mathbf{r}}(a) = 0$ whenever well-defined, and the event has probability 0. Assume a is not the only outgoing edge from v. Consider any fixed $\mathbf{r}_{-a} \in \mathbb{R}^{E\setminus\{a\}}$. We bound the probability of the event

$$\mathcal{E} \triangleq \left\{\mathbf{r}(a) \in \mathbb{R} : \mu^{\mathrm{mp},\mathbf{r}}(v) = \mu^{\mathrm{mp},\mathbf{r}}(v') \text{ and } 0 < \left|\Delta^{1,\mathbf{r}}(a)\right| < \delta\right\}.$$

Formally, we prove $\mathbb{P}_{\xi_a}[\mathcal{E}] \le 2\delta\phi$.

Let $\mu^{\gamma,\mathbf{r}_{-a}}$ denote the optimal value vector of the restricted game played on the edge set $E\setminus\{a\}$. Since this restricted game is completely independent of edge a, its value vector $\mu^{\gamma,\mathbf{r}_{-a}}$ is a function depending only on $\mathbf{r}_{-a}$. For any $x \in \mathcal{E}$, let $\mathbf{r} = (x, \mathbf{r}_{-a})$. Then, we have $\Delta^{1,\mathbf{r}}(a) \neq 0$. This implies that the edge a is strictly suboptimal under the Blackwell optimal strategies. Consequently, for γ sufficiently close to 1, the optimal strategy does not select a, meaning the value vectors of the full game and the restricted game exactly match: $\mu^{\gamma,\mathbf{r}} = \mu^{\gamma,\mathbf{r}_{-a}}$. Taking the limit $\gamma \uparrow 1$ of $(1-\gamma)\mu^{\gamma,\mathbf{r}}$ and applying Eq. (2) gives $\mu^{\mathrm{mp},\mathbf{r}_{-a}}(u) = \mu^{\mathrm{mp},\mathbf{r}}(u)$ for all $u \in V$. In particular, the assumption $\mu^{\mathrm{mp},\mathbf{r}}(v) = \mu^{\mathrm{mp},\mathbf{r}}(v')$ implies $\mu^{\mathrm{mp},\mathbf{r}_{-a}}(v) = \mu^{\mathrm{mp},\mathbf{r}_{-a}}(v')$.

Under this condition, the limit of the suboptimality gap evaluates to:

$$\Delta^{1,\mathbf{r}}(a) = x + \lim_{\gamma\uparrow 1}\left(\gamma\mu^{\gamma,\mathbf{r}_{-a}}(v') - \mu^{\gamma,\mathbf{r}_{-a}}(v)\right).$$

Let $C(\mathbf{r}_{-a})$ denote the limit term. Applying the value decomposition (Lemma 2) to the restricted game and using $\mu^{\mathrm{mp},\mathbf{r}_{-a}}(v) = \mu^{\mathrm{mp},\mathbf{r}_{-a}}(v')$, the singular $\frac{1}{1-\gamma}$ terms cancel, so $C(\mathbf{r}_{-a})$ exists and is finite. Since $C(\mathbf{r}_{-a})$ is fixed with respect to x, the condition $0 < \left|x + C(\mathbf{r}_{-a})\right| < \delta$ requires the value of x to fall strictly within the interval $\left(-C(\mathbf{r}_{-a}) - \delta, -C(\mathbf{r}_{-a}) + \delta\right)$. Let $\lambda(\mathcal{E})$ denote the Lebesgue measure of the set of x values satisfying this condition. Note that $\lambda(\mathcal{E}) \le 2\delta$. Since the probability density of x is uniformly bounded by ϕ, the probability of the event satisfies $\mathbb{P}_{\xi_a}[\mathcal{E}] \le 2\delta\phi$. Taking expectation over the independent randomness in ξ_{-a} gives

$$\mathbb{P}_\xi[\mathcal{E}] = \mathop{\mathbb{E}}_{\xi_{-a}}\left[\mathbb{P}_{\xi_a}[\mathcal{E}]\right] \le 2\delta\phi,$$

which completes the proof. □

C.3 Ingredient 3: Mean-payoff Value Separation

Lemma 5 (Mean-payoff separation)**.** *Let G be any game graph and let $v, v' \in V$. Under the smoothed model, for any $B > 0$, we have*

$$\mathbb{P}_\xi\Big[0 < \big|\mu^{mp,\mathbf{r}}(v) - \mu^{mp,\mathbf{r}}(v')\big| < B\Big] \le 2Bmn\phi.$$

To prove the above result, we first define some useful notation. We then establish a reachability separation lemma, which is a deterministic statement about the structure of Blackwell optimal strategies. Finally, we use this lemma to prove the mean-payoff value separation result.

Cycles. Let $\mathcal{C}$ denote the set of simple directed cycles of G. For a cycle $C \in \mathcal{C}$, let $E(C)$ and $V(C)$ be its edge set and vertex set, and write

$$\mathrm{avg}_{\mathbf{r}}(C) \triangleq \frac{1}{|E(C)|} \sum_{a \in E(C)} \mathbf{r}(a).$$

Reachable vertices and edges. For a strategy $\pi_{\min} \in \Pi_{\min}$ of the Min Player, let $V_{\pi_{\min}}(v)$ be the set of vertices reachable from v after fixing Min to $\pi_{\min}$ and leaving all Max edges available. Let $E_{\pi_{\min}}(v)$ be the set of edges reachable from v in this one-player game, i.e.,

$$E_{\pi_{\min}}(v) \triangleq \big\{(v', v'') \in E : v' \in V_{\pi_{\min}}(v) \cap V_{\max}\big\} \cup \big\{(v', \pi_{\min}(v')) : v' \in V_{\pi_{\min}}(v) \cap V_{\min}\big\}.$$

Symmetrically, for a strategy $\pi_{\max} \in \Pi_{\max}$ of the Max Player, let $V_{\pi_{\max}}(v)$ be the set of vertices reachable from v after fixing Max to $\pi_{\max}$ and leaving all Min edges available, and let

$$E_{\pi_{\max}}(v) \triangleq \big\{(v', v'') \in E : v' \in V_{\pi_{\max}}(v) \cap V_{\min}\big\} \cup \big\{(v', \pi_{\max}(v')) : v' \in V_{\pi_{\max}}(v) \cap V_{\max}\big\}.$$

For a strategy $\pi_{\min} \in \Pi_{\min}$, the cycles $C \in \mathcal{C}$ with $E(C) \subseteq E_{\pi_{\min}}(v)$ are exactly the simple directed cycles reachable from v after fixing $\pi_{\min}$, and likewise for $\pi_{\max} \in \Pi_{\max}$.

Lemma 15 (Reachability separation)**.** *Let $\mathbf{r} \in \mathbb{R}^E$, let $(\pi^{bw}_{\min}, \pi^{bw}_{\max})$ be a Blackwell optimal strategy pair, and let $v, v' \in V$. The cycle $C^{\mathbf{r}}_v$ is the recurrent cycle reached from v under $(\pi^{bw}_{\min}, \pi^{bw}_{\max})$.*

1. *If $\mu^{mp,\mathbf{r}}(v) > \mu^{mp,\mathbf{r}}(v')$, then*
$$E(C^{\mathbf{r}}_v) \cap E_{\pi^{bw}_{\min}}(v') = \varnothing,$$
2. *If $\mu^{mp,\mathbf{r}}(v) < \mu^{mp,\mathbf{r}}(v')$, then*
$$E(C^{\mathbf{r}}_v) \cap E_{\pi^{bw}_{\max}}(v') = \varnothing.$$

Proof. Since the pair $(\pi^{\mathrm{bw}}_{\min}, \pi^{\mathrm{bw}}_{\max})$ is Blackwell optimal, it is also mean-payoff optimal [BK76]. Therefore, $\pi^{\mathrm{bw}}_{\min}$ is an optimal min strategy and $\pi^{\mathrm{bw}}_{\max}$ an optimal max strategy for the game $G^{\mathrm{mp},\mathbf{r}}$. Suppose $E(C^{\mathbf{r}}_v) \cap E_{\pi^{\mathrm{bw}}_{\min}}(v') \neq \varnothing$. Then, there exists a vertex $u \in C^{\mathbf{r}}_v \cap V_{\pi^{\mathrm{bw}}_{\min}}(v')$. Since $C^{\mathbf{r}}_v$ follows $\pi^{\mathrm{bw}}_{\min}$ at every Min vertex, traversing it from u gives $E(C^{\mathbf{r}}_v) \subseteq E_{\pi^{\mathrm{bw}}_{\min}}(v')$, so some Max strategy $\pi_{\max}$ makes $\rho(v', \pi^{\mathrm{bw}}_{\min}, \pi_{\max})$ reach $C^{\mathbf{r}}_v$ and cycle it forever, giving $u^{\mathrm{mp},\mathbf{r}}(\rho(v', \pi^{\mathrm{bw}}_{\min}, \pi_{\max})) = \mathrm{avg}_{\mathbf{r}}(C^{\mathbf{r}}_v)$. As $\pi^{\mathrm{bw}}_{\min}$ is an optimal min strategy, this payoff is at most $\mu^{\mathrm{mp},\mathbf{r}}(v')$, and therefore

$$\mu^{\mathrm{mp},\mathbf{r}}(v) = \mathrm{avg}_{\mathbf{r}}(C^{\mathbf{r}}_v) \le \mu^{\mathrm{mp},\mathbf{r}}(v'),$$

contradicting $\mu^{\mathrm{mp},\mathbf{r}}(v) > \mu^{\mathrm{mp},\mathbf{r}}(v')$. Hence $E(C^{\mathbf{r}}_v) \cap E_{\pi^{\mathrm{bw}}_{\min}}(v') = \varnothing$.

The proof of the second statement is symmetric. Suppose $E(C^{\mathbf{r}}_v) \cap E_{\pi^{\mathrm{bw}}_{\max}}(v') \neq \varnothing$. Then, there exists a vertex $u \in C^{\mathbf{r}}_v \cap V_{\pi^{\mathrm{bw}}_{\max}}(v')$. Since $C^{\mathbf{r}}_v$ follows $\pi^{\mathrm{bw}}_{\max}$ at every Max vertex, traversing it from u gives $E(C^{\mathbf{r}}_v) \subseteq E_{\pi^{\mathrm{bw}}_{\max}}(v')$, so some Min strategy $\pi_{\min}$ makes $\rho(v', \pi_{\min}, \pi^{\mathrm{bw}}_{\max})$ reach $C^{\mathbf{r}}_v$ and cycle it forever, giving $u^{\mathrm{mp},\mathbf{r}}(\rho(v', \pi_{\min}, \pi^{\mathrm{bw}}_{\max})) = \mathrm{avg}_{\mathbf{r}}(C^{\mathbf{r}}_v)$. As $\pi^{\mathrm{bw}}_{\max}$ is an optimal max strategy, this payoff is at least $\mu^{\mathrm{mp},\mathbf{r}}(v')$, and therefore

$$\mu^{\mathrm{mp},\mathbf{r}}(v) = \mathrm{avg}_{\mathbf{r}}(C^{\mathbf{r}}_v) \ge \mu^{\mathrm{mp},\mathbf{r}}(v'),$$

contradicting $\mu^{\mathrm{mp},\mathbf{r}}(v) < \mu^{\mathrm{mp},\mathbf{r}}(v')$. Hence $E(C^{\mathbf{r}}_v) \cap E_{\pi^{\mathrm{bw}}_{\max}}(v') = \varnothing$. □

Proof of Lemma 5. If $v = v'$, the event is empty; assume $v \neq v'$. Let

$$\mathcal{E}_{\mathrm{bw}} \triangleq \big\{\mathbf{r} \in \mathbb{R}^E : \exists!\, (\pi^{\mathrm{bw}}_{\min}, \pi^{\mathrm{bw}}_{\max}) \in \Pi_{\min} \times \Pi_{\max} \text{ that is Blackwell optimal for } \mathbf{r}\big\}.$$

By Corollary 13, $\mathbb{P}_\xi[\mathcal{E}_{\mathrm{bw}}] = 1$, so it suffices to bound the two bad events

$$\mathcal{B}^+ \triangleq \mathcal{E}_{\mathrm{bw}} \cap \big\{\mathbf{r} : 0 < \mu^{\mathrm{mp},\mathbf{r}}(v) - \mu^{\mathrm{mp},\mathbf{r}}(v') < B\big\},$$
$$\mathcal{B}^- \triangleq \mathcal{E}_{\mathrm{bw}} \cap \big\{\mathbf{r} : 0 < \mu^{\mathrm{mp},\mathbf{r}}(v') - \mu^{\mathrm{mp},\mathbf{r}}(v) < B\big\}.$$

We bound $\mathbb{P}_\xi[\mathcal{B}^+]$; the bound on $\mathbb{P}_\xi[\mathcal{B}^-]$ is identical after exchanging v and v'.

Given an edge $a \in E$ and rewards $\mathbf{r}_{-a}$, when $(x; \mathbf{r}_{-a}) \in \mathcal{E}_{\mathrm{bw}}$ let $\pi^{\mathrm{bw}}_{\min}$ be its Blackwell optimal Min strategy and $C_v^{(x;\mathbf{r}_{-a})}$ the recurrent cycle reached from v, and define the slice

$$S_a^+(\mathbf{r}_{-a}) \triangleq \big\{x \in \mathbb{R} : (x; \mathbf{r}_{-a}) \in \mathcal{E}_{\mathrm{bw}},\ 0 < \mu^{\mathrm{mp},(x;\mathbf{r}_{-a})}(v) - \mu^{\mathrm{mp},(x;\mathbf{r}_{-a})}(v') < B,\ a \in E(C_v^{(x;\mathbf{r}_{-a})}),\ a \notin E_{\pi^{\mathrm{bw}}_{\min}}(v')\big\}.$$

For rewards $\mathbf{r} \in \mathcal{B}^+$, let $C_v^{\mathbf{r}}$ be the recurrent cycle reached from v under its unique Blackwell optimal pair. Therefore, we have $\mathrm{avg}_{\mathbf{r}}(C_v^{\mathbf{r}}) = \mu^{\mathrm{mp},\mathbf{r}}(v)$. By Lemma 15, for all edges $a \in E(C_v^{\mathbf{r}})$, we have $a \notin E_{\pi^{\mathrm{bw}}_{\min}}(v')$. Thus, $\mathbf{r}(a) \in S_a^+(\mathbf{r}_{-a})$. Consequently, we have

$$\mathcal{B}^+ \subseteq \bigcup_{a \in E} \big\{\mathbf{r} : \mathbf{r}(a) \in S_a^+(\mathbf{r}_{-a})\big\}.$$

Moreover, we show that $\mu^{\mathrm{mp},(x;\mathbf{r}_{-a})}(v')$ is constant on $S_a^+(\mathbf{r}_{-a})$. For a fixed Min strategy $\pi_{\min}$, every play from v' is a lasso whose recurrent cycle C satisfies $E(C) \subseteq E_{\pi_{\min}}(v')$, and the Max Player can reach and repeat any such cycle; hence the largest payoff the Max Player secures from v' against $\pi_{\min}$ is $\max\{\mathrm{avg}_{(x;\mathbf{r}_{-a})}(C) : C \in \mathcal{C},\ E(C) \subseteq E_{\pi_{\min}}(v')\}$. Since the mean-payoff value is the min-max of the payoff,

$$\mu^{\mathrm{mp},(x;\mathbf{r}_{-a})}(v') = \min_{\pi_{\min} \in \Pi_{\min}} \max_{\substack{C \in \mathcal{C} \\ E(C) \subseteq E_{\pi_{\min}}(v')}} \mathrm{avg}_{(x;\mathbf{r}_{-a})}(C).$$

For $x \in S_a^+(\mathbf{r}_{-a})$, the Blackwell optimal $\pi^{\mathrm{bw}}_{\min}$ is an optimal min strategy, so it attains this minimum, and it satisfies $a \notin E_{\pi^{\mathrm{bw}}_{\min}}(v')$; hence the minimum is attained among the strategies with $a \notin E_{\pi_{\min}}(v')$:

$$\mu^{\mathrm{mp},(x;\mathbf{r}_{-a})}(v') = \min_{\substack{\pi_{\min} \in \Pi_{\min} \\ a \notin E_{\pi_{\min}}(v')}} \max_{\substack{C \in \mathcal{C} \\ E(C) \subseteq E_{\pi_{\min}}(v')}} \mathrm{avg}_{(x;\mathbf{r}_{-a})}(C).$$

Every cycle on the right-hand side avoids a, so its average does not involve $x = \mathbf{r}(a)$; hence $\mu^{\mathrm{mp},(x;\mathbf{r}_{-a})}(v')$ is a constant α depending only on $\mathbf{r}_{-a}$ for all $x \in S_a^+(\mathbf{r}_{-a})$. Let $f(x) \triangleq \mu^{\mathrm{mp},(x;\mathbf{r}_{-a})}(v)$. The function f is continuous, piecewise affine, nondecreasing, and absolutely continuous because it is the min-max over finitely many positional strategy pairs, and for each such pair the recurrent-cycle mean is an affine function of x with slope 0 or $1/c$ for some $c \leq n$. Let

$$D \triangleq \bigcup_{\substack{C, C' \in \mathcal{C} \\ \mathrm{avg}_{(\cdot;\mathbf{r}_{-a})}(C) \not\equiv \mathrm{avg}_{(\cdot;\mathbf{r}_{-a})}(C')}} \big\{x \in \mathbb{R} : \mathrm{avg}_{(x;\mathbf{r}_{-a})}(C) = \mathrm{avg}_{(x;\mathbf{r}_{-a})}(C')\big\},$$

which is finite and, since each affine piece of f coincides with $\mathrm{avg}_{(x;\mathbf{r}_{-a})}(C)$ for some $C \in \mathcal{C}$, contains every breakpoint of f. Fix a connected component J of $\mathbb{R} \setminus D$ with $J \cap S_a^+(\mathbf{r}_{-a}) \neq \varnothing$. Because every breakpoint of f belongs to D, the restriction of f to J consists of a single affine piece. Hence there exists a cycle $\widehat{C} \in \mathcal{C}$ such that

$$f(x) = \mathrm{avg}_{(x;\mathbf{r}_{-a})}(\widehat{C}) \qquad \text{for every } x \in J.$$

Choose $x_0 \in J \cap S_a^+(\mathbf{r}_{-a})$ and let

$$C_0 \triangleq C_v^{(x_0;\mathbf{r}_{-a})}.$$

By the definition of $S_a^+(\mathbf{r}_{-a})$, we have $(x_0;\mathbf{r}_{-a}) \in \mathcal{E}_{\text{bw}}$ and $a \in E(C_0)$. The Blackwell optimal strategy pair at $(x_0;\mathbf{r}_{-a})$ induces the recurrent cycle C_0 from v, and therefore

$$f(x_0) = \text{avg}_{(x_0;\mathbf{r}_{-a})}(C_0) = \text{avg}_{(x_0;\mathbf{r}_{-a})}(\widehat{C}).$$

If the two affine functions $\text{avg}_{(\cdot;\mathbf{r}_{-a})}(C_0)$ and $\text{avg}_{(\cdot;\mathbf{r}_{-a})}(\widehat{C})$ were not identical, their equality at x_0 would imply $x_0 \in D$ by the definition of D, contradicting $x_0 \in J \subseteq \mathbb{R} \setminus D$. Consequently, these affine functions are identical, and hence

$$f(x) = \text{avg}_{(x;\mathbf{r}_{-a})}(C_0) \qquad \text{for every } x \in J.$$

The recurrent cycle C_0 is simple, contains a, and has at most n edges. Thus, the variable reward $x = \mathbf{r}(a)$ occurs exactly once in the sum defining its cycle average, and therefore

$$f'(x) = \frac{1}{|E(C_0)|} \geq \frac{1}{n} \qquad \text{for all } x \in J.$$

Every $x \in S_a^+(\mathbf{r}_{-a}) \setminus D$ lies in such a component, so $f'(x) \geq 1/n$ for almost every $x \in S_a^+(\mathbf{r}_{-a})$. Since $\mu^{\text{mp},(x;\mathbf{r}_{-a})}(v') = \alpha$ on $S_a^+(\mathbf{r}_{-a})$, we have $S_a^+(\mathbf{r}_{-a}) \subseteq f^{-1}\big((\alpha, \alpha+B)\big)$. Therefore, since f is nondecreasing and absolutely continuous, we have

$$\lambda\big(S_a^+(\mathbf{r}_{-a})\big) \leq n \int_{S_a^+(\mathbf{r}_{-a})} f'(x)\,dx \leq n \int_{f^{-1}((\alpha,\alpha+B))} f'(x)\,dx \leq nB,$$

where $\lambda(\cdot)$ denotes the Lebesgue measure.

Conditioning on $\mathbf{r}_{-a}$, the reward $\mathbf{r}(a) = \mathbf{r}_0(a) + \xi_a$ has density at most ϕ and is independent of $\mathbf{r}_{-a}$, so

$$\mathbb{P}_\xi[\mathcal{B}^+] \leq \sum_{a\in E} \mathbb{P}_\xi\big[\mathbf{r}(a) \in S_a^+(\mathbf{r}_{-a})\big] \leq \sum_{a\in E} \phi \mathop{\mathbb{E}}_{\mathbf{r}_{-a}}\big[\lambda\big(S_a^+(\mathbf{r}_{-a})\big)\big] \leq mnB\phi.$$

The identical argument gives $\mathbb{P}_\xi[\mathcal{B}^-] \leq mnB\phi$, and therefore

$$\mathbb{P}_\xi\Big[0 < \big|\mu^{\text{mp},\mathbf{r}}(v) - \mu^{\text{mp},\mathbf{r}}(v')\big| < B\Big] \leq \mathbb{P}_\xi[\mathcal{B}^+] + \mathbb{P}_\xi[\mathcal{B}^-] \leq 2mnB\phi.$$

□

D Proofs of Section 7

In this section, we provide the proofs omitted from Section 7, organized by the same structure as the main body.

D.1 Tail Bound

Lemma 6 (Tail bound)**.** *Let G be any game graph. Under the smoothed model, for any $x \geq 1$, with the bounding function defined as $M(x) \triangleq 1 + \frac{1}{\theta}\ln(mx)$, we have*

$$\mathbb{P}_\xi\big[\gamma_{bw}(\mathbf{r}) > 1 - \tfrac{1}{x}\big] \leq \frac{KM(x)}{x}.$$

Proof. For any $M \geq 1$, $B > 0$, and $\delta > 0$, define the events

$$\begin{aligned}
\mathcal{E}_1(M) &\triangleq \big\{\mathbf{r} \in \mathbb{R}^E : \|\mathbf{r}\|_\infty \leq M\big\},\\
\mathcal{E}_2(B) &\triangleq \big\{\mathbf{r} \in \mathbb{R}^E : \big|\mu^{\text{mp},\mathbf{r}}(v) - \mu^{\text{mp},\mathbf{r}}(v')\big| \in \{0\} \cup [B,\infty) \text{ for all } (v,v') \in E\big\},\\
\mathcal{E}_3(\delta) &\triangleq \big\{\mathbf{r} \in \mathbb{R}^E : \big|\Delta^{1,\mathbf{r}}(a)\big| \in \{0\} \cup [\delta,\infty) \text{ for all } a=(v,v') \in E \text{ with } \mu^{\text{mp},\mathbf{r}}(v) = \mu^{\text{mp},\mathbf{r}}(v')\big\},\\
\mathcal{E}_4 &\triangleq \big\{\mathbf{r} \in \mathbb{R}^E : \forall v \in V,\ \exists!\, a=(v,v') \in E \text{ such that } \mu^{\text{mp},\mathbf{r}}(v) = \mu^{\text{mp},\mathbf{r}}(v') \text{ and } \Delta^{1,\mathbf{r}}(a) = 0\big\}.
\end{aligned}$$

Let $L \triangleq 18n^2M$. Conditioned on $\mathcal{E}_1(M) \cap \mathcal{E}_2(B) \cap \mathcal{E}_3(\delta) \cap \mathcal{E}_4$, we show that $\gamma_{\mathrm{bw}}(\mathbf{r}) \le \max\big(0, 1 - \frac{\delta}{2L}, 1 - \frac{B}{10nM}\big) = \max\big(0, 1 - \frac{\delta}{36n^2M}, 1 - \frac{B}{10nM}\big)$.

Indeed, consider $\mathbf{r} \in \mathcal{E}_1(M) \cap \mathcal{E}_2(B) \cap \mathcal{E}_3(\delta) \cap \mathcal{E}_4$. Since $\mathbf{r} \in \mathcal{E}_4$, by Proposition 12, for each vertex $v \in V$ there exists a unique edge $a^\star(v) = (v, v') \in E$ with $\mu^{\mathrm{mp},\mathbf{r}}(v) = \mu^{\mathrm{mp},\mathbf{r}}(v')$ and $\Delta^{1,\mathbf{r}}(a^\star(v)) = 0$. Therefore, by Corollary 13, there exists a unique Blackwell optimal strategy pair $(\pi^{\mathrm{bw}}_{\min}, \pi^{\mathrm{bw}}_{\max})$, which selects $a^\star(v)$ at each vertex $v \in V$.

By Lemma 3 and Lemma 2, since $\mathbf{r} \in \mathcal{E}_1(M)$, the L-Lipschitz bound on the suboptimality gap and the value decomposition both hold on $(\gamma_{\mathrm{bw}}(\mathbf{r}), 1)$. We argue by contradiction: suppose $\gamma_{\mathrm{bw}}(\mathbf{r}) > \max\big(1 - \frac{\delta}{2L}, 1 - \frac{B}{10nM}\big)$. Then for any $\gamma \in (\gamma_{\mathrm{bw}}(\mathbf{r}), 1)$, $1 - \gamma < 1 - \gamma_{\mathrm{bw}}(\mathbf{r}) < \min\big(\frac{\delta}{2L}, \frac{B}{10nM}\big)$. We bound $\big|\Delta^{\gamma,\mathbf{r}}(a)\big|$ for any non-optimal edge $a = (v, v') \neq a^\star(v)$ on $(\gamma_{\mathrm{bw}}(\mathbf{r}), 1)$, splitting into two cases.

Case $\mu^{\mathrm{mp},\mathbf{r}}(v) = \mu^{\mathrm{mp},\mathbf{r}}(v')$. Since $\mathbf{r} \in \mathcal{E}_3(\delta)$ and a is non-optimal, $\big|\Delta^{1,\mathbf{r}}(a)\big| \ge \delta$. Then:

$$\begin{aligned}\big|\Delta^{\gamma,\mathbf{r}}(a)\big| &\ge \big|\Delta^{1,\mathbf{r}}(a)\big| - \big|\Delta^{\gamma,\mathbf{r}}(a) - \Delta^{1,\mathbf{r}}(a)\big| \\ &\ge \delta - L(1-\gamma) > \frac{\delta}{2} > 0,\end{aligned}$$

where the first inequality is the triangle inequality, the second combines $\big|\Delta^{1,\mathbf{r}}(a)\big| \ge \delta$ with the L-Lipschitz bound on $(\gamma_{\mathrm{bw}}(\mathbf{r}), 1)$, and the strict third inequality uses $1 - \gamma < \frac{\delta}{2L}$.

Case $\mu^{\mathrm{mp},\mathbf{r}}(v) \neq \mu^{\mathrm{mp},\mathbf{r}}(v')$. Since $\mathbf{r} \in \mathcal{E}_2(B)$, $\big|\mu^{\mathrm{mp},\mathbf{r}}(v') - \mu^{\mathrm{mp},\mathbf{r}}(v)\big| \ge B$. Substituting the value decomposition from Lemma 2 into the suboptimality gap yields

$$\Delta^{\gamma,\mathbf{r}}(a) = \mathbf{r}(v, v') - \mu^{\mathrm{mp},\mathbf{r}}(v') + \frac{\mu^{\mathrm{mp},\mathbf{r}}(v') - \mu^{\mathrm{mp},\mathbf{r}}(v)}{1-\gamma} + \gamma w^{\mathbf{r}}_{v'}(\gamma) - w^{\mathbf{r}}_{v}(\gamma).$$

Applying the reverse triangle inequality to isolate the singular term:

$$\begin{aligned}|\Delta^{\gamma,\mathbf{r}}(a)| &\ge \frac{|\mu^{\mathrm{mp},\mathbf{r}}(v') - \mu^{\mathrm{mp},\mathbf{r}}(v)|}{1-\gamma} - \big|\mathbf{r}(v, v') - \mu^{\mathrm{mp},\mathbf{r}}(v') + \gamma w^{\mathbf{r}}_{v'}(\gamma) - w^{\mathbf{r}}_{v}(\gamma)\big| \\ &\ge \frac{B}{1-\gamma} - \big(2M + 4nM + 4nM\big) \\ &\ge \frac{B}{1-\gamma} - 10nM > 0,\end{aligned}$$

where the first inequality is the reverse triangle inequality together with $\big|\mu^{\mathrm{mp},\mathbf{r}}(v') - \mu^{\mathrm{mp},\mathbf{r}}(v)\big| \ge B$, the second uses $|\mathbf{r}(v, v')| \le M$ and $|\mu^{\mathrm{mp},\mathbf{r}}(v')| \le M$ (from $\mathcal{E}_1(M)$) together with $|w^{\mathbf{r}}_{v}(\gamma)|, |w^{\mathbf{r}}_{v'}(\gamma)| \le 4nM$ from Lemma 2, the third uses $n \ge 1$, and the final strict inequality uses $1 - \gamma < \frac{B}{10nM}$.

Combining the two cases, $\big|\Delta^{\gamma,\mathbf{r}}(a)\big| > 0$ for every non-optimal edge a and every $\gamma \in (\gamma_{\mathrm{bw}}(\mathbf{r}), 1)$. By definition of Blackwell optimality, on $(\gamma_{\mathrm{bw}}(\mathbf{r}), 1)$ the pair $(\pi^{\mathrm{bw}}_{\min}, \pi^{\mathrm{bw}}_{\max})$ is optimal, so for non-optimal edges a outgoing from v, the sign of $\Delta^{\gamma,\mathbf{r}}(a)$ is positive if $v \in V_{\min}$ and negative if $v \in V_{\max}$. Since $\Delta^{\gamma,\mathbf{r}}(a)$ is continuous in γ on $[0, 1)$ and its magnitude stays strictly positive throughout $(\gamma_{\mathrm{bw}}(\mathbf{r}), 1)$, continuity extends the strictly positive magnitude and the consistent sign to a left-neighborhood $(\gamma_{\mathrm{bw}}(\mathbf{r}) - \eta, \gamma_{\mathrm{bw}}(\mathbf{r})]$ for some $\eta > 0$ uniform over the finite edge set. On this neighborhood, $a^\star(v)$ is the unique edge at v achieving $\Delta^{\gamma,\mathbf{r}}(a^\star(v)) = 0$, so the pair $(\pi^{\mathrm{bw}}_{\min}, \pi^{\mathrm{bw}}_{\max})$ remains optimal on $(\gamma_{\mathrm{bw}}(\mathbf{r}) - \eta, 1)$. This contradicts the definition of $\gamma_{\mathrm{bw}}(\mathbf{r})$ as the infimum of valid thresholds. Therefore, $\gamma_{\mathrm{bw}}(\mathbf{r}) \le \max\big(1 - \frac{\delta}{36n^2M}, 1 - \frac{B}{10nM}\big)$.

For a given $x \ge 1$, we substitute the designated $M(x) = 1 + \frac{1}{\theta}\ln(mx)$, $\delta(x) \triangleq \frac{36n^2M(x)}{x}$, and $B(x) \triangleq \frac{10nM(x)}{x}$. By definition, $\frac{\delta(x)}{36n^2M(x)} = \frac{B(x)}{10nM(x)} = \frac{1}{x}$, which implies $\gamma_{\mathrm{bw}}(\mathbf{r}) \le \max\big(0, 1 - \frac{1}{x}\big)$. Under the event $\mathcal{E}_1(M(x)) \cap \mathcal{E}_2(B(x)) \cap \mathcal{E}_3(\delta(x)) \cap \mathcal{E}_4$, we have $\gamma_{\mathrm{bw}}(\mathbf{r}) \le 1 - \frac{1}{x}$. Therefore, the event $\gamma_{\mathrm{bw}}(\mathbf{r}) > 1 - \frac{1}{x}$ is upper bounded:

$$\mathbb{P}_\xi\big[\gamma_{\mathrm{bw}}(\mathbf{r}) > 1 - \tfrac{1}{x}\big] \le \mathbb{P}_\xi\big[\overline{\mathcal{E}_1}(M(x))\big] + \mathbb{P}_\xi\big[\overline{\mathcal{E}_2}(B(x))\big] + \mathbb{P}_\xi\big[\overline{\mathcal{E}_3}(\delta(x))\big] + \mathbb{P}_\xi\big[\overline{\mathcal{E}_4}\big].$$

For $\mathbb{P}_\xi\big[\overline{\mathcal{E}_1}(M(x))\big]$, we have

$$\begin{aligned}\mathbb{P}_\xi\big[\overline{\mathcal{E}_1}(M(x))\big] &\le m\exp\big(-\theta(M(x)-1)\big)\\ &= m\exp\big(-\theta\frac{1}{\theta}\ln(mx)\big) = \frac{1}{x},\end{aligned}$$

where the first inequality is due to Eq. (1) and a union bound over the m edges, and the first equality is due to the definition of $M(x)$. For $\mathbb{P}_\xi\big[\overline{\mathcal{E}_2}(B(x))\big]$, we have

$$\begin{aligned}\mathbb{P}_\xi\big[\overline{\mathcal{E}_2}(B(x))\big] &\le 2B(x)m^2n\phi\\ &= 2m^2n\frac{10nM(x)}{x}\phi = \frac{20m^2n^2\phi M(x)}{x},\end{aligned}$$

where the first inequality is due to Lemma 5 and a union bound over the m edges, and the first equality is due to the definition of $B(x)$. For $\mathbb{P}_\xi\big[\overline{\mathcal{E}_3}(\delta(x))\big]$, we have

$$\begin{aligned}\mathbb{P}_\xi\big[\overline{\mathcal{E}_3}(\delta(x))\big] &\le 2m\delta(x)\phi\\ &= 2m\frac{36n^2M(x)}{x}\phi = \frac{72mn^2\phi M(x)}{x},\end{aligned}$$

where the first inequality is due to Lemma 4 and a union bound over the m edges, and the first equality is due to the definition of $\delta(x)$. We also have $\mathbb{P}_\xi\big[\overline{\mathcal{E}_4}\big] = 0$ by Proposition 12. Therefore, we obtain

$$\begin{aligned}\mathbb{P}_\xi\big[\gamma_{\mathrm{bw}}(\mathbf{r}) > 1-\tfrac{1}{x}\big] &\le \frac{1}{x} + \frac{20m^2n^2\phi M(x)}{x} + \frac{72mn^2\phi M(x)}{x}\\ &= \frac{1+4mn^2(5m+18)\phi M(x)}{x}\\ &\le \frac{1+92m^2n^2\phi M(x)}{x} \le \frac{KM(x)}{x},\end{aligned}$$

where the second inequality uses $5m+18 \le 23m$ (which holds since $m \ge 1$), and the last inequality uses $1 \le 92m^2n^2M(x)$ (which holds since $M(x) \ge 1$ for $x \ge 1$ and $m, n \ge 1$) and the definition of K in Eq. (3). This completes the proof. □

D.2 Restarting Policy Iteration

Lemma 7 (Total iteration bound)**.** *Let G be any game graph with $n \ge 2$, let $\gamma \in [0,1)$, and let $\mathbf{r} \in \mathbb{R}^E$. Write $N_{iter}(\mathbf{r})$ for the total number of iterations executed by $\mathsf{RePI}(G,\gamma,\mathbf{r})$ (Algorithm 2), summed over all calls to PI. Then $N_{iter}(\mathbf{r}) \le (\lceil n\ln n\rceil + 2)\,n^n$ for every $\mathbf{r}$, and, under the smoothed model, with probability 1,*

$$N_{iter}(\mathbf{r}) \le 27\cdot\frac{m}{1-\gamma_{bw}(\mathbf{r})}\ln\Big(\frac{e\cdot n}{1-\gamma_{bw}(\mathbf{r})}\Big).$$

Proof. Let $k_{\max} = \lceil n\ln n\rceil$ and let $k_{\text{final}} \in \{0,1,\dots,k_{\max}\}$ denote the index of the last iteration of the for-loop executed by RePI. The algorithm performs PI at threshold γ_k for each $k \in \{0,1,\dots,k_{\text{final}}\}$ (or at the target γ when $\gamma \le \gamma_{k_{\text{final}}}$), and optionally one fallback $\mathsf{PI}(G,\gamma,\mathbf{r})$ in Phase 3. By Proposition 1, each PI call at threshold $\gamma_k = 1-e^{-k}$ executes at most $\frac{6m}{1-\gamma_k}\ln\big(\frac{n}{1-\gamma_k}\big) = 6m\,e^k(\ln n + k)$ iterations, and is also trivially bounded by n^n because the number of iterations is upper bounded by the number of positional strategies. Therefore, we have

$$N_{\text{iter}}(\mathbf{r}) \le \sum_{k=0}^{k_{\text{final}}} \min\big(6m\,e^k(\ln n + k),\, n^n\big) + n^n\cdot\mathbf{1}\{\text{Phase 3 triggered}\}.$$

Deterministic bound. Bounding every term in the sum by n^n and using that there are at most $k_{\max}+1$ terms plus at most one fallback gives $N_{\text{iter}}(\mathbf{r}) \le (\lceil n \ln n \rceil + 2)\, n^n$ for every $\mathbf{r}$.

Smoothed geometric bound. Let $\Gamma_{\text{alg}} \triangleq \{\gamma_k : k = 0, 1, \ldots, k_{\max}\}$ be the finite set of thresholds used by the algorithm. By Lemma 14, with probability 1, for every $\gamma_k \in \Gamma_{\text{alg}}$ with $\gamma_k > \gamma_{\text{bw}}(\mathbf{r})$, the game $G^{\gamma_k,\mathbf{r}}$ has a unique optimal strategy pair, and this pair is Blackwell optimal. Condition on this probability-one event. We bound the total in terms of $\gamma_{\text{bw}}(\mathbf{r})$, splitting into two cases according to whether the Blackwell threshold is resolved within the for-loop. In each case we establish the two bounds

$$e^{k_{\text{final}}+1} \le \frac{e^2}{1-\gamma_{\text{bw}}(\mathbf{r})} \qquad \text{and} \qquad \ln n + k_{\text{final}} \le \ln\Big(\frac{en}{1-\gamma_{\text{bw}}(\mathbf{r})}\Big). \tag{4}$$

Case 1: $\gamma_{\text{bw}}(\mathbf{r}) < \gamma_{k_{\max}}$. Since $\gamma_0 = 0$, there exists $i \in \{0, 1, \ldots, k_{\max}-1\}$ with $\gamma_i \le \gamma_{\text{bw}}(\mathbf{r}) < \gamma_{i+1}$. Because $\gamma_{i+1} > \gamma_{\text{bw}}(\mathbf{r})$, the conditioned event implies that the unique optimal pair of $G^{\gamma_{i+1},\mathbf{r}}$ is Blackwell optimal, so $\mathsf{PI}(G, \gamma_{i+1}, \mathbf{r})$ returns a Blackwell optimal pair; being Blackwell optimal, it is optimal for every $\gamma' \in [\gamma_{\text{bw}}(\mathbf{r}), 1)$. Hence, if iteration $k = i+1$ reaches Phase 2 (that is, if $\gamma > \gamma_{i+1}$), the returned pair is optimal for the target γ and Phase 2 succeeds; otherwise $\gamma \le \gamma_{i+1}$ and the algorithm has already returned $\mathsf{PI}(G, \gamma, \mathbf{r})$ at some iteration $k \le i+1$. Either way the algorithm terminates without triggering Phase 3, so $k_{\text{final}} \le i+1$ and the Phase 3 term vanishes. From $1-\gamma_i = e^{-i} \ge 1-\gamma_{\text{bw}}(\mathbf{r})$ we obtain $e^i \le 1/(1-\gamma_{\text{bw}}(\mathbf{r}))$, hence

$$e^{k_{\text{final}}+1} \le e^{i+2} \le \frac{e^2}{1-\gamma_{\text{bw}}(\mathbf{r})} \qquad \text{and} \qquad k_{\text{final}} \le \ln \frac{1}{1-\gamma_{\text{bw}}(\mathbf{r})} + 1,$$

which gives Eq. (4) since $\ln n + k_{\text{final}} \le \ln n + \ln \frac{1}{1-\gamma_{\text{bw}}(\mathbf{r})} + 1 = \ln\big(en/(1-\gamma_{\text{bw}}(\mathbf{r}))\big)$.

Case 2: $\gamma_{\text{bw}}(\mathbf{r}) \ge \gamma_{k_{\max}}$. Here $1/(1-\gamma_{\text{bw}}(\mathbf{r})) \ge e^{k_{\max}} \ge n^n$, and the algorithm may run all $k_{\max}+1$ for-loop iterations plus a Phase 3 fallback, so $k_{\text{final}} = k_{\max}$. Hence

$$e^{k_{\text{final}}+1} = e \cdot e^{k_{\max}} \le \frac{e}{1-\gamma_{\text{bw}}(\mathbf{r})} \le \frac{e^2}{1-\gamma_{\text{bw}}(\mathbf{r})} \qquad \text{and} \qquad k_{\text{final}} = k_{\max} \le \ln \frac{1}{1-\gamma_{\text{bw}}(\mathbf{r})},$$

which gives Eq. (4) since $\ln n + k_{\text{final}} \le \ln n + \ln \frac{1}{1-\gamma_{\text{bw}}(\mathbf{r})} \le \ln\big(en/(1-\gamma_{\text{bw}}(\mathbf{r}))\big)$.

We now bound the for-loop contribution using Eq. (4):

$$\begin{aligned}
\sum_{k=0}^{k_{\text{final}}} 6m\, e^k (\ln n + k) &\le 6m\, (\ln n + k_{\text{final}}) \sum_{k=0}^{k_{\text{final}}} e^k \\
&\le 6m\, (\ln n + k_{\text{final}}) \cdot \frac{e^{k_{\text{final}}+1}}{e-1} \\
&\le \frac{6e^2}{e-1} \cdot \frac{m}{1-\gamma_{\text{bw}}(\mathbf{r})} \ln\Big(\frac{en}{1-\gamma_{\text{bw}}(\mathbf{r})}\Big),
\end{aligned}$$

where the first inequality uses $\ln n + k \le \ln n + k_{\text{final}}$ for every $k \le k_{\text{final}}$, the second sums the geometric series $\sum_{k=0}^{k_{\text{final}}} e^k \le e^{k_{\text{final}}+1}/(e-1)$, and the third substitutes both bounds of Eq. (4). It remains to account for the Phase 3 fallback. In Case 1 this term vanishes; in Case 2 it contributes at most

$$n^n \le \frac{1}{1-\gamma_{\text{bw}}(\mathbf{r})} \le \frac{m}{1-\gamma_{\text{bw}}(\mathbf{r})} \ln\Big(\frac{en}{1-\gamma_{\text{bw}}(\mathbf{r})}\Big),$$

using $m \ge 1$ and $\ln(en) \ge 1$. Adding the two contributions yields

$$N_{\text{iter}}(\mathbf{r}) \le \Big(\frac{6e^2}{e-1} + 1\Big) \cdot \frac{m}{1-\gamma_{\text{bw}}(\mathbf{r})} \ln\Big(\frac{en}{1-\gamma_{\text{bw}}(\mathbf{r})}\Big) \le 27 \cdot \frac{m}{1-\gamma_{\text{bw}}(\mathbf{r})} \ln\Big(\frac{en}{1-\gamma_{\text{bw}}(\mathbf{r})}\Big).$$

which is the second bound. As it was derived on the probability-one event from Lemma 14, this bound holds with probability 1. □

D.3 High Probability Guarantees

Lemma 16. *Let G be any game graph. Under the smoothed model, for any $\epsilon \in (0,1)$, define*

$$L_\epsilon \triangleq \theta + \ln\Big(1 + \frac{2mK}{\epsilon\theta}\Big) \quad \text{and} \quad x_\epsilon \triangleq \frac{2KL_\epsilon}{\epsilon\theta}.$$

With probability at least $1-\epsilon$ over the perturbation $\mathbf{r}$, Algorithm 2 executes PI *for a total of at most $27\, m\, x_\epsilon \ln(en\, x_\epsilon)$ iterations. In particular, x_ϵ is polynomial in $n, m, \phi, 1/\theta$, and $1/\epsilon$.*

Proof. Since $L_\epsilon \geq \theta$, $K \geq 1$, $m \geq 1$, and $\epsilon \in (0,1)$, we have $x_\epsilon \geq \frac{2K}{\epsilon} > 2$.

We claim that $\gamma_{\mathrm{bw}}(\mathbf{r}) \leq 1 - 1/x_\epsilon$ with probability at least $1-\epsilon$. Indeed, applying Lemma 6 with $x = x_\epsilon$ yields

$$\mathbb{P}_\xi\big[\gamma_{\mathrm{bw}}(\mathbf{r}) > 1 - \tfrac{1}{x_\epsilon}\big] \leq \frac{KM(x_\epsilon)}{x_\epsilon} = \frac{\epsilon}{2L_\epsilon}\Big(\theta + \ln(mx_\epsilon)\Big),$$

where the equality uses the definitions of $M(x)$ and x_ϵ. By the definition of L_ϵ,

$$\ln(mx_\epsilon) = \ln\Big(\frac{2mK}{\epsilon\theta}\Big) + \ln L_\epsilon \leq \ln\Big(1 + \frac{2mK}{\epsilon\theta}\Big) + L_\epsilon = 2L_\epsilon - \theta,$$

where the inequality uses $\ln L_\epsilon \leq L_\epsilon$. Consequently,

$$\frac{KM(x_\epsilon)}{x_\epsilon} \leq \frac{\epsilon}{2L_\epsilon}(\theta + 2L_\epsilon - \theta) = \epsilon,$$

which yields $\gamma_{\mathrm{bw}}(\mathbf{r}) \leq 1 - 1/x_\epsilon$ with probability at least $1-\epsilon$. As the geometric bound of Lemma 7 holds with probability 1, its intersection with this event still has probability at least $1-\epsilon$; assume both hold for the rest of the proof.

The map $y \mapsto y\ln(en\, y)$ is increasing on $y \geq 1$. Therefore, we have

$$\frac{1}{1-\gamma_{\mathrm{bw}}(\mathbf{r})} \ln\Big(\frac{en}{1-\gamma_{\mathrm{bw}}(\mathbf{r})}\Big) \leq x_\epsilon \ln(en\, x_\epsilon).$$

Lemma 7 then yields $N_{\mathrm{iter}}(\mathbf{r}) \leq 27\, m\, x_\epsilon \ln(en\, x_\epsilon)$. □

D.4 Expected Runtime Guarantees

Lemma 9 (Bounding expected truncated inverse gap)**.** *Let G be any game graph with $n \geq 2$. Let K be defined in Eq. (3). Under the smoothed model, we have*

$$\mathbb{E}_\xi\big[Y(\mathbf{r})\big] \leq \frac{6(\theta+1)}{\theta} K (n \ln m)^3.$$

Proof. Let x_0 be the unique real number satisfying $x_0 \geq 1$ and $x_0 \ln x_0 = n^n$. We compute the expectation of $Y(\mathbf{r})$ by integrating the tail probability and applying the substitution $y = x\ln x$, where $dy = (1+\ln x)dx$:

$$\begin{aligned}
\mathbb{E}_\xi\big[Y(\mathbf{r})\big] &= \int_0^{n^n} \mathbb{P}_\xi\big[Y(\mathbf{r}) > y\big]\, dy \\
&= \int_1^{x_0} \mathbb{P}_\xi\big[\gamma_{\mathrm{bw}}(\mathbf{r}) > 1 - \tfrac{1}{x}\big](1+\ln x)\, dx \\
&\leq \int_1^{e} 1 \cdot (1+\ln x)\, dx + \int_e^{x_0} \mathbb{P}_\xi\big[\gamma_{\mathrm{bw}}(\mathbf{r}) > 1 - \tfrac{1}{x}\big](1+\ln x)\, dx \\
&= e + \int_e^{x_0} \mathbb{P}_\xi\big[\gamma_{\mathrm{bw}}(\mathbf{r}) > 1 - \tfrac{1}{x}\big](1+\ln x)\, dx.
\end{aligned} \tag{5}$$

For the second integral, we substitute the bound $\mathbb{P}_\xi[\gamma_{\mathrm{bw}}(\mathbf{r}) > 1 - \frac{1}{x}] \le \frac{KM(x)}{x}$ established in Lemma 6, which applies on $[e, x_0]$ since $x \ge e > 1$:

$$\int_e^{x_0} \mathbb{P}_\xi\big[\gamma_{\mathrm{bw}}(\mathbf{r}) > 1 - \tfrac{1}{x}\big](1 + \ln x)\, dx \le K \int_e^{x_0} \frac{M(x)(1+\ln x)}{x}\, dx.$$

We apply the substitution $u = \ln x$, meaning $dx/x = du$. The integration limits shift to $u \in [1, \ln x_0]$. Since $x_0 \le n^n$, the upper limit is strictly bounded by $n \ln n$. This yields:

$$\begin{aligned}
\int_e^{x_0} \frac{M(x)(1+\ln x)}{x}\, dx &\le \int_1^{n\ln n} M(e^u)(1+u)\, du \\
&\le M(e^{n\ln n}) \int_1^{n\ln n} (1+u)\, du \\
&= \Big(1 + \frac{\ln m + n\ln n}{\theta}\Big)\big[u + \frac{u^2}{2}\big]_1^{n\ln n} \\
&\le \Big(1 + \frac{\ln m + n\ln n}{\theta}\Big)\big(\frac{1}{2}(n\ln n)^2 + n\ln n\big) \\
&\le \Big(1 + \frac{\ln m + n\ln n}{\theta}\Big)\frac{3}{2}(n\ln n)^2 \\
&\le \frac{\theta+1}{\theta}(\ln m + n\ln n)\frac{3}{2}(n\ln n)^2 \\
&\le \frac{\theta+1}{\theta}\, 2n\ln m\, \frac{3}{2}(n\ln n)^2 \\
&\le \frac{\theta+1}{\theta}\, 2n\ln m\, \frac{3}{2}(n\ln m)^2 \\
&= \frac{3(\theta+1)}{\theta}(n\ln m)^3,
\end{aligned}$$

where the first equality follows from $M(e^u) = 1 + \frac{\ln m + u}{\theta}$; the inequality $\frac{1}{2}(n\ln n)^2 + n\ln n \le \frac{3}{2}(n\ln n)^2$ uses $n\ln n \le (n\ln n)^2$ (which holds for all $n \ge 1$, since $n\ln n \ge 0$); the bound $1 + \frac{\ln m + n\ln n}{\theta} \le \frac{\theta+1}{\theta}(\ln m + n\ln n)$ uses $1 \le \ln m + n\ln n$ (which holds for $n \ge 2$); the bounds $\ln m + n\ln n \le 2n\ln m$ and $(n\ln n)^2 \le (n\ln m)^2$ both use $m \ge n$.

Substituting this bound back into Eq. (5) yields $\mathbb{E}_\xi[Y(\mathbf{r})] \le e + K\frac{3(\theta+1)}{\theta}(n\ln m)^3$. Finally, since $\frac{\theta+1}{\theta} \ge 1$ and $n\ln m \ge 2\ln 2 > 1$ (as $m \ge n \ge 2$), we have $\frac{3(\theta+1)}{\theta}(n\ln m)^3 \ge 3(n\ln m)^3 > 3 > e$; since $K \ge 1$, this gives $e \le K\frac{3(\theta+1)}{\theta}(n\ln m)^3$, and hence $\mathbb{E}_\xi[Y(\mathbf{r})] \le 2K\frac{3(\theta+1)}{\theta}(n\ln m)^3$, completing the proof. □

Lemma 17. *Let G be any game graph with $n \ge 2$. Under the smoothed model, the expected total number of iterations in all the* PI*'s executed by Algorithm 2 is bounded by*

$$324\, m\ln n\, K\frac{3(\theta+1)}{\theta}(n\ln m)^3,$$

where K is defined in Eq. (3). In particular, this bound is polynomial in n, m, ϕ, and $1/\theta$.

Proof. Write $x \triangleq 1/(1 - \gamma_{\mathrm{bw}}(\mathbf{r})) \ge 1$, and recall the truncated inverse gap $Y(\mathbf{r}) = \min\big(x\ln x,\, n^n\big)$. By Lemma 7, with probability 1 we have

$$N_{\mathrm{iter}}(\mathbf{r}) \le \min\Big(27\, m\, x\ln(en\, x),\ (\lceil n\ln n\rceil + 2)\, n^n\Big).$$

We claim that

$$N_{\mathrm{iter}}(\mathbf{r}) \le 27\, m\big((\ln n + 2)\, Y(\mathbf{r}) + e(\ln n + 1)\big). \tag{6}$$

Indeed, we prove this by splitting into two cases according to the truncation.

Case 1: $x \ln x \le n^n$**.** Here $Y(\mathbf{r}) = x \ln x$, and we bound the minimum by its first term:

$$\begin{aligned} N_{\text{iter}}(\mathbf{r}) &\le 27\, m\, x \ln(en\, x) \\ &= 27\, m\big(x \ln x + (\ln n + 1)x\big) \\ &\le 27\, m\big(Y(\mathbf{r}) + (\ln n + 1)(Y(\mathbf{r}) + e)\big) \\ &= 27\, m\big((\ln n + 2)Y(\mathbf{r}) + e(\ln n + 1)\big), \end{aligned}$$

where the first equality uses $\ln(en\, x) = \ln x + \ln n + 1$, and the second inequality uses $x \ln x = Y(\mathbf{r})$ together with $x \le x \ln x + e = Y(\mathbf{r}) + e$, valid for all $x \ge 1$.

Case 2: $x \ln x > n^n$**.** Here $Y(\mathbf{r}) = n^n$, and we bound the minimum by its second term. Since every vertex has at least one outgoing edge, $m \ge n$, hence $\lceil n \ln n \rceil + 2 \le n \ln n + 3 \le 27\, m(\ln n + 2)$. Therefore

$$N_{\text{iter}}(\mathbf{r}) \le (\lceil n \ln n \rceil + 2)\, n^n \le 27\, m(\ln n + 2)\, n^n \le 27\, m\big((\ln n + 2)Y(\mathbf{r}) + e(\ln n + 1)\big).$$

Consequently, Eq. (6) holds with probability 1; since $N_{\text{iter}}(\mathbf{r}) \le (\lceil n \ln n \rceil + 2)\, n^n$ for every $\mathbf{r}$ by Lemma 7, the complementary null event contributes nothing to $\mathbb{E}_\xi[N_{\text{iter}}(\mathbf{r})]$. Taking the expectation of Eq. (6) over the smoothed perturbation $\mathbf{r}$ and applying Lemma 9,

$$\begin{aligned} \mathbb{E}_\xi[N_{\text{iter}}(\mathbf{r})] &\le 27\, m\big((\ln n + 2)\, \mathbb{E}_\xi[Y(\mathbf{r})] + e(\ln n + 1)\big) \\ &\le 27\, m\Big(2(\ln n + 2)K \frac{3(\theta+1)}{\theta} (n \ln m)^3 + e(\ln n + 1)\Big) \\ &\le 81\, m(\ln n + 2)K \frac{3(\theta+1)}{\theta} (n \ln m)^3 \\ &\le 324\, m \ln n\, K \frac{3(\theta+1)}{\theta} (n \ln m)^3, \end{aligned}$$

where the third inequality uses $e(\ln n + 1) \le (\ln n + 2)K \frac{3(\theta+1)}{\theta}(n \ln m)^3$, which follows from $e \le K \frac{3(\theta+1)}{\theta}(n \ln m)^3$ (since $K \ge 1$ and $\frac{3(\theta+1)}{\theta}(n \ln m)^3 \ge 3(n \ln m)^3 \ge 3 > e$), and the last inequality uses $\ln n + 2 \le 4 \ln n$ (which holds for $n \ge 2$). This is the claimed bound. □

E Mean-payoff Games

In this section, we provide the technical development for mean-payoff games that establishes Corollary 11, combining a high-probability guarantee (Lemma 19) and an expected runtime guarantee (Lemma 20). We first adapt the restarting algorithm of Section 7.2 to the mean-payoff setting and establish a total iteration bound for it (Section E.1). Combining this bound with the tail bound of Section 7.1 yields smoothed polynomial runtime with high probability (Section E.2); combining it with the expected truncated inverse gap of Section 7.4 yields smoothed polynomial expected runtime (Section E.3).

E.1 Restarting Mean-payoff Policy Iteration

In this subsection, we adapt the restarting variant of policy iteration to the mean-payoff setting, called $\mathsf{RePI}_{\text{mp}}$, presented in full detail in Algorithm 3, and then establish a deterministic bound on its total iteration count in terms of the Blackwell threshold $\gamma_{\text{bw}}(\mathbf{r})$ (Lemma 18).

Informal description. Our algorithm sweeps through the same geometric sequence of thresholds $\gamma_k = 1 - e^{-k}$ for $k = 0, 1, \ldots, \lceil n \ln n \rceil$, with each successive threshold e times closer to 1 than the previous one. For each k in turn, it solves the game with the discount factor γ_k via the standard discounted-sum $\mathsf{PI}(G, \gamma_k, \mathbf{r})$ to obtain a candidate strategy pair $(\pi_{\min}, \pi_{\max})$. The algorithm then

tests whether this candidate pair is already optimal for the mean-payoff game. For this purpose, we use the combinatorial strategy evaluation of [BV07], which reduces mean-payoff games to the longest shortest paths problem and evaluates strategies using shortest distances to a sink. Their method can determine whether a given positional strategy pair is mean-payoff optimal; we invoke this test as a black box. If the test succeeds, the algorithm computes the mean-payoff value vector induced by the candidate pair and returns both. Otherwise it proceeds to the next threshold γ_{k+1} and retries. If every threshold in the sequence fails to be optimal, the algorithm falls back to the longest-shortest-path strategy-improvement algorithm of [BV07], denoted by $\mathsf{PI}_{\mathrm{mp}}(G, \mathbf{r})$. We use this algorithm as a deterministic black-box fallback: it returns a mean-payoff optimal positional strategy pair and value vector, and for our purposes the upper bound of at most n^n strategy improvements is sufficient, which is due to the fact that there are at most n^n positional strategies pairs in a game with n vertices.

Algorithm 3 Restarting Mean-payoff Policy Iteration: $\mathsf{RePI}_{\mathrm{mp}}(G, \mathbf{r})$

Input: Graph $G = (V_{\min}, V_{\max}, E)$ with $n = |V| \geq 2$, reward vector $\mathbf{r} \in \mathbb{R}^E$.
Output: Mean-payoff optimal strategy pair $(\pi^*_{\min}, \pi^*_{\max})$ and mean-payoff value vector $\mu^{\mathrm{mp},\mathbf{r}}$ for the game $G^{\mathrm{mp},\mathbf{r}}$.
Let $k_{\max} = \lceil n \ln n \rceil$.
for $k = 0, 1, \ldots, k_{\max}$ **do**
    Set the threshold discount factor $\gamma_k = 1 - e^{-k}$.
    // **Phase 1: Solve the discounted game at the threshold**
    $(\pi_{\min}, \pi_{\max}, \mu_{\gamma_k}) \leftarrow \mathsf{PI}(G, \gamma_k, \mathbf{r})$
    // **Phase 2: Mean-payoff optimality test**
    Use the longest shortest paths algorithm of [BV07] to test whether $(\pi_{\min}, \pi_{\max})$ is mean-payoff optimal.
    if $(\pi_{\min}, \pi_{\max})$ is mean-payoff optimal **then**
        Compute the mean-payoff value vector $\mu^{\mathrm{mp},\mathbf{r}}$ induced by $(\pi_{\min}, \pi_{\max})$.
        return $(\pi_{\min}, \pi_{\max}, \mu^{\mathrm{mp},\mathbf{r}})$
    end if
end for
// **Phase 3: Fallback**
return $\mathsf{PI}_{\mathrm{mp}}(G, \mathbf{r})$

The intuition mirrors the discounted-sum case. Under the smoothed model, with high probability some threshold γ_k in the sequence strictly exceeds the Blackwell threshold $\gamma_{\mathrm{bw}}(\mathbf{r})$. Thus, the strategy pair returned by $\mathsf{PI}(G, \gamma_k, \mathbf{r})$ is Blackwell optimal and consequently mean-payoff optimal [BK76], so the Phase 2 test succeeds and the algorithm terminates without reaching the fallback.

Lemma 18 (Total iteration bound)**.** *Let G be any game graph with $n \geq 2$ and let $\mathbf{r} \in \mathbb{R}^E$; write $N_{iter}(\mathbf{r})$ for the total number of iterations performed by $\mathsf{RePI}_{mp}(G, \mathbf{r})$ (Algorithm 3), across all calls to PI and the optional PI_{mp} fallback. Then $N_{iter}(\mathbf{r}) \leq (\lceil n \ln n \rceil + 2)\, n^n$ for every $\mathbf{r}$, and, under the smoothed model, with probability 1,*

$$N_{iter}(\mathbf{r}) \leq 27 \cdot \frac{m}{1 - \gamma_{bw}(\mathbf{r})} \ln\Big(\frac{en}{1 - \gamma_{bw}(\mathbf{r})}\Big).$$

Proof. The deterministic bound $N_{\mathrm{iter}}(\mathbf{r}) \leq (\lceil n \ln n \rceil + 2)\, n^n$ holds exactly as in Lemma 7: there are at most $\lceil n \ln n \rceil + 2$ calls, each terminating in at most n^n iterations (the fallback $\mathsf{PI}_{\mathrm{mp}}(G, \mathbf{r})$ likewise, by [BV07]). For the geometric bound, condition on the probability-one event from Lemma 14 with

$\Gamma_{\mathrm{alg}} = \{\gamma_k : k = 0, 1, \ldots, k_{\max}\}$, and run the same case analysis as in Lemma 7 with two minor changes. First, when $\gamma_{i+1} > \gamma_{\mathrm{bw}}(\mathbf{r})$, the unique optimal pair of $G^{\gamma_{i+1},\mathbf{r}}$ is Blackwell optimal and is also *mean-payoff* optimal [BK76], so $\mathsf{PI}(G, \gamma_{i+1}, \mathbf{r})$ returns it; the test based on longest shortest paths from [BV07] therefore certifies the pair as mean-payoff optimal in Phase 2, and the algorithm terminates without triggering Phase 3. Second, when Phase 3 is triggered (i.e., $\gamma_{\mathrm{bw}}(\mathbf{r}) \geq \gamma_{k_{\max}}$), the fallback is $\mathsf{PI}_{\mathrm{mp}}(G, \mathbf{r})$ rather than $\mathsf{PI}(G, \gamma, \mathbf{r})$; it terminates in at most n^n iterations [BV07], and the same bound $n^n \leq \frac{m}{1-\gamma_{\mathrm{bw}}(\mathbf{r})} \ln\!\big(\frac{en}{1-\gamma_{\mathrm{bw}}(\mathbf{r})}\big)$ applies in this regime. The geometric sum then combines as in Lemma 7 to give the second bound, which holds with probability 1 by Lemma 14. □

E.2 High Probability Guarantees

In this subsection, we combine the total iteration bound of Lemma 18 with the tail bound from Section 7.1 to obtain a high-probability bound on the number of iterations of $\mathsf{RePI}_{\mathrm{mp}}$.

Lemma 19. *Let G be any game graph. Under the smoothed model, for any $\epsilon \in (0, 1)$, let x_ϵ be as defined in Lemma 16. With probability at least $1 - \epsilon$ over the perturbation $\mathbf{r}$, RePI_{mp} (Algorithm 3) executes a total of at most $27\, m\, x_\epsilon \ln(en\, x_\epsilon)$ iterations.*

Proof. By Step 1 of the proof of Lemma 16, $\gamma_{\mathrm{bw}}(\mathbf{r}) \leq 1 - 1/x_\epsilon$ with probability at least $1-\epsilon$. As the geometric bound of Lemma 18 holds with probability 1, assume this event together with it (their intersection still has probability at least $1 - \epsilon$), so $\frac{1}{1-\gamma_{\mathrm{bw}}(\mathbf{r})} \leq x_\epsilon$, and the map $y \mapsto y \ln(en\, y)$ being increasing on $y \geq 1$ gives $\frac{1}{1-\gamma_{\mathrm{bw}}(\mathbf{r})} \ln\!\big(\frac{en}{1-\gamma_{\mathrm{bw}}(\mathbf{r})}\big) \leq x_\epsilon \ln(en\, x_\epsilon)$. Lemma 18 then yields $N_{\mathrm{iter}}(\mathbf{r}) \leq 27\, m\, x_\epsilon \ln(en\, x_\epsilon)$. □

E.3 Expected Runtime Guarantees

In this subsection, we combine the total iteration bound of Lemma 18 with the bound on the expected truncated inverse gap from Lemma 9 to obtain a bound on the expected number of iterations of $\mathsf{RePI}_{\mathrm{mp}}$.

Lemma 20. *Let G be any game graph. Under the smoothed model, the expected total number of iterations executed by Algorithm 3 (across all PI executions and the fallback PI_{mp}, when triggered) is bounded by*

$$324\, m \ln n\, K \frac{3(\theta + 1)}{\theta} (n \ln m)^3,$$

where K is defined in Eq. (3). In particular, this bound is polynomial in n, m, ϕ, and $1/\theta$.

Proof. The argument is identical to the proof of Lemma 17, invoking the total iteration bound of Lemma 18 in place of Lemma 7. With probability 1 the same pointwise bound $N_{\mathrm{iter}}(\mathbf{r}) \leq 27\, m\big((\ln n + 2)Y(\mathbf{r}) + e(\ln n + 1)\big)$ holds, and since $N_{\mathrm{iter}}(\mathbf{r}) \leq (\lceil n \ln n \rceil + 2)n^n$ for every $\mathbf{r}$ the complementary null event contributes nothing to the expectation; taking the expectation and applying Lemma 9 yields the stated bound. □

Proof of Corollary 11. The high-probability guarantee follows from Lemma 19. The expected runtime guarantee follows from Lemma 20. □