Asymmetrically Discounted Stochastic Games

Sarvin Bahmani
University of Liverpool, UK

Soumyajit Paul
University of Liverpool, UK

Sven Schewe
University of Liverpool, UK

Shadi Tasdighi Kalat
University of Colorado Boulder, CO, USA

Ashutosh Trivedi
University of Colorado Boulder, CO, USA
University of Liverpool, UK

Abstract

We study asymmetrically discounted stochastic games, in which players use distinct and reasonably apart discount factors. We show that optimal strategies in these games may require both memory and randomisation, in contrast to the classical symmetrically discounted setting. Our main technical contribution establishes that computing incentive Stackelberg equilibria – a variant of Stackelberg equilibria in which one player, called Player Max, can offer payments to the other player, called Player Min – is no harder than solving classical discounted games. We further show that optimal strategies in this setting can be realised by finite counting strategies, whereas restricting players to stationary strategies makes the problem computationally intractable. Finally, we establish that computing classical Stackelberg equilibria in these games under the constraint of memoryless strategies is NP-complete and remains NP-hard even when general or counting strategies are allowed.

2012 ACM Subject Classification Mathematics of computing → Markov processes; Theory of computation → Convergence and learning in games

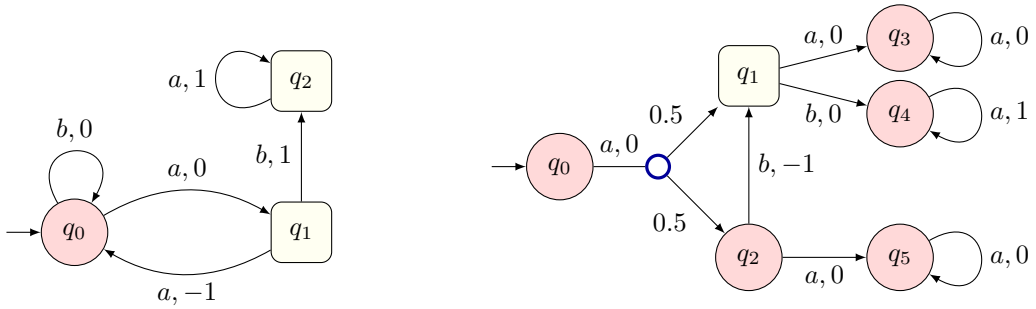
Keywords and phrases Stochastic Games, Asymmetric Discounting, Stackelberg Equilibrium

Digital Object Identifier 10.4230/LIPIcs.CONCUR.2026.14

Funding This work was supported by the EPSRC through grants EP/X03688X/1 (TRUSTED) and EP/X042596/1 (Games for Good), and in part by the NSF under CAREER Award CCF-2146563. Ashutosh Trivedi is a Royal Society Wolfson Visiting Fellow and gratefully acknowledges the support of the Wolfson Foundation and the Royal Society.

1 Introduction

Since their introduction by Shapley in 1953, *stochastic games* [26] have served as a foundational framework for modelling adversarial interactions between two rational agents. In these games, two players, commonly referred to as Player Min, or simply Min, and Player Max, or simply Max, interact on a finite-state game arena by taking actions that move a token from one state to another. Because the interaction is zero-sum, each action is associated with a payoff: a gain for one player corresponds to an equal loss for the other. The game proceeds indefinitely, but with a constant probability of termination at each step, typically modelled by a geometric distribution. This termination behaviour is formalised through a discount factor, which serves a dual purpose: it bounds the expected duration of play and gives the game desirable mathematical properties, including contraction and determinacy [26].



(a) Lower value less than upper value. **(b)** Players need both memory and chance.

Figure 1 Illustrations of ASGs. Circles denote states controlled by Player Min, and boxes denote states controlled by Player Max.

Beyond its technical role, the discount factor also captures a notion of *time preference*: the idea that a dollar today is worth more than a dollar tomorrow. While classical stochastic games assume symmetric time preferences for the two players, we study rational decision-making in settings where agents have distinct discount factors. We call these games *asymmetrically discounted stochastic games* (ASGs).

Asymmetry in Time Preference. Time preference refers to the tendency of agents to favour rewards that arrive sooner over rewards that are delayed. Different agents, however, need not share the same time preference: two agents may agree that immediate rewards are preferable to delayed ones, yet differ substantially in how strongly they discount the future. The study of time preference and discounting spans a wide range of disciplines, including psychology [23], economics [25, 1], game theory [20], control theory, and optimisation [24]. In psychology, time discounting has been closely linked to self-control and delayed gratification, most notably through Mischel's influential *Marshmallow experiments* [23]. These studies suggest that individuals with a higher effective discount factor, who are willing to wait longer for a larger reward, tend to exhibit traits associated with better long-term outcomes. Subsequent work [4] has examined additional factors influencing time preferences, including personal characteristics and contextual factors.

Put simply, while most agents agree that \$1 today is preferable to \$1 tomorrow, they differ in how much future rewards must compensate for delay. For instance, one agent may prefer \$2 tomorrow to \$1 today, whereas another may be willing to wait only if the delayed reward is at least \$3. In formal models of decision-making, such variation in time preference is captured through discounting [28]. Given this variability, it is unlikely that two rational agents interacting in a stochastic game arena would share exactly the same discount factor. This observation makes the classical symmetric-discounting assumption fragile in practice. In particular, we show that as long as the discount factors differ, optimal strategies may require both memory and randomisation.

Non-determinacy in Asymmetric Games. A class of games is said to be *determined* if the optimal payoff of the players is invariant under rational play when the order in which the players announce their strategies is changed. Equivalently, the optimal payoff that Max can guarantee when he announces his strategy first, allowing Min to respond optimally, coincides with the optimal payoff that Max can obtain when he plays second, after observing Min's strategy. Here, the latter assumes that Min plays optimally with respect to her own payoff, rather than adversarially minimising Max's payoff.

Consider the game depicted in Figure 1a, with discount factors $\alpha = 0.9$ for Max and $\beta = 0.1$ for Min. The optimal payoff for Max when he reveals his strategy second is 0. Indeed, Min controls only state q_0, and her best strategy when announcing first is to remain in q_0 by playing action b. This yields payoff 0 for both Max and Min. If Min instead takes action a, then Max responds optimally by choosing b from q_1, which is worse for Min. Hence, 0 is the best payoff Max can obtain when he goes second.

On the other hand, when Max goes first, he can use a randomised strategy to obtain a different payoff. In particular, consider a strategy for Max in which he transitions to q_0 with probability p and to q_2 with probability $1 - p$. By choosing $p = \frac{10}{19}$, Max can ensure a payoff of $\frac{720}{109} \approx 6.6$, while Min's expected payoff remains 0 when she chooses a rather than b at q_0. Thus, Min's payoff does not improve by choosing b over a at state q_0. Recall that Min's objective is to optimise her payoff, not to hurt Max by minimising Max's payoff.

Suppose Max increases p slightly beyond $\frac{10}{19}$, say to $p' = \frac{10}{19} + \epsilon$ for some small $\epsilon > 0$. This shift makes the expected discounted payoff of choosing edge (q_1, q_0) strictly higher than that of choosing edge (q_1, q_2) for Min, thereby changing her optimal action. Consequently, Max's ability to influence Min through stochastic choice creates a discontinuity in value: although Min can restrict Max's payoff to 0 when Max goes second, Max can exploit the asymmetry in time preferences to secure a higher payoff using mixed strategies when he announces his strategy upfront. This demonstrates that the payoff values differ depending on the order in which the players announce their strategies, and hence the game is *non-determined*. More details on the example in Figure 1a can be found in Appendix A.

▶ **Theorem 1.** *Asymmetrically discounted stochastic games are not determined.*

Proof. The example in Figure 1a shows that the upper and lower values of the game can be distinct. Hence, asymmetrically discounted stochastic games are not determined. ◀

ASGs need not have a "minimax" value: Max may fail to achieve the upper value using any positional strategy, yet can surpass the lower value using mixed strategies. This separation between the lower and upper values, arising from asymmetric discounting and the necessity of mixing, motivates the need to explore alternative notions of solution stability. In this context, we investigate variants of *Stackelberg equilibria* [29] as solution concepts for ASGs.

The need for memory. To further illustrate the challenges in solving ASGs, we show that Max may need memory to play optimally. Consider the game in Figure 1b, with Min's discount factor $\beta = \frac{2}{3}$ and Max's discount factor $\alpha = 0.9$. At state q_1, Max must choose differently depending on the path by which the state is reached. If q_1 is reached directly from q_0, then Max can choose action b, thereby maximising his payoff. By contrast, if q_1 is reached from q_2, then Max must mix between his actions in order to make it unprofitable for Min to choose action a from q_2 instead of action b. The required mixing probability depends on Min's discount factor.

If Max randomises at q_1 by choosing action b with probability p and action a with probability $1 - p$, then p should be chosen such that actions a and b yield the same payoff for Min at q_2. Consider the following strategy profile with strategy σ for Max and τ for Min:

$$\begin{aligned} \sigma(q_0, a, q_1) &= b, \quad \text{and} \quad \sigma(q_0, a, q_2, b, q_1) = (1-p) : a \ + \ p : b \quad , \\ \tau(q_0, a, q_2) &= b. \end{aligned}$$

Notice that the payoff for Max using this strategy is

$$0.5 \left(\frac{\alpha^2}{1-\alpha} \right) + 0.5 \left(-\alpha + \frac{p\alpha^3}{1-\alpha} \right),$$

The payoff for Min, on the other hand, is

$$0.5\left(\frac{\beta^2}{1-\beta}\right) + 0.5\left(-\beta + \frac{p\beta^3}{1-\beta}\right).$$

Note that, at state q_2, Min is deciding between the values $0.5\left(-\beta + \frac{p\beta^3}{1-\beta}\right)$ and 0. For the strategy profile to be stable, these two values must coincide, yielding $p = \frac{1-\beta}{\beta^2} = \frac{3}{4}$. For Max, it returns the maximum payoff among all stable profiles which, for $\alpha = 0.9$, is 6.334. Min's payoff is $\frac{2}{3}$ regardless of the choice. Consequently, Min has no profitable deviation and follows the proposed strategy, choosing action b at state q_2.

Strategy profiles in which the follower does not benefit from deviating are called Stackelberg equilibria (SE) [29]. In our setting, play begins with the leader proposing a profile. The follower can accept or reject this proposed profile only *before the first move*; if the follower accepts, the players enter into a contract that both subsequently follow. Thus, the follower's comparison point is the payoff she would obtain in the two-player discounted stochastic game under her discount factor β. She accepts the proposed profile unless this comparison value is better for her. These equilibria capture the strategic advantage of commitment and allow the leader to influence the follower's choices through foresight and control of information.

Stackelberg and Incentive Stackelberg Equilibria. In non-zero-sum games, Nash equilibria are a standard solution concept. However, they treat the players symmetrically and therefore restrict the strategy profiles that Max can induce. For this reason, we focus on SE, which explicitly model asymmetric commitment power and are more favourable to Max.

We assume that the player with the larger discount factor, Max, acts as the leader and can credibly commit to a strategy. The follower, Min, observes this commitment and responds optimally to maximise her own payoff. Since a higher discount factor corresponds to a longer effective planning horizon, it is natural in our setting to designate Max as the leader.

In the game shown in Figure 1a, with discount factors 0.9 for Max and 0.1 for Min, we saw earlier that the Stackelberg equilibrium value for Max is 6.6. We then consider a refined solution concept in which Max can additionally offer an explicit incentive, in the form of a reward, to induce Min to follow a prescribed strategy profile. In the same game, if Max offers a reward of $\frac{1}{9}$ and proposes the strategy profile in which Min plays a at q_0 and Max plays b at q_1, then Max obtains a payoff of $9 - \frac{1}{9} \approx 8.89$, while Min receives 0.

Under this mechanism, the proposed profile is optimal for Min, and Max achieves a higher payoff than under classical Stackelberg equilibrium, since Min optimises her own payoff rather than minimising Max's. Consequently, deviating from the proposed profile does not improve Min's payoff. We refer to this solution concept as an *Incentive Stackelberg Equilibrium* (ISE), the primary focus of this paper. Such refinements of Stackelberg equilibria, where one player can offer incentives to another, have been studied in [13, 15, 14] and provide a more robust solution concept in settings with a designated leader.

Contributions. We initiate the study of two-player asymmetrically discounted stochastic games (ASGs) on graphs with zero-sum rewards on actions. Our primary focus is on ISE, a refinement of classical Stackelberg equilibria. Our main contributions are as follows:

- We show that ASGs are not determined, and that optimal equilibria may require both memory and randomisation, even under Stackelberg equilibria.

- From a computational perspective, we study the restriction to memoryless strategies. We show that computing incentive equilibria under this restriction is NP-complete. We further establish NP-completeness for ordinary Stackelberg equilibria with stationary strategies, and show that this hardness extends to counting strategies.
- As our main technical contribution, we present an algorithm computing optimal incentive Stackelberg equilibria. We prove that, when discount factors are reasonably apart, this computation is no harder than solving classical discounted stochastic games. Moreover, the resulting equilibrium strategies are simple, requiring only finite counting memory.
- We implement our algorithm and evaluate it on random games and PRISM-games benchmarks, demonstrating practical scalability and illustrating how performance depends on the separation between the players' discount factors.

Related Work. Asymmetric discounting has been studied in repeated games, where Fink [21] showed that positional Nash equilibria exists. Lehrer and Pauzner [20] showed that differing discount factors expand the equilibrium payoff set, since more patient players can delay gratification to support cooperation, effectively subsidising impatient players early on. Dasgupta and Ghosh [10] further analysed how heterogeneity in time preferences reshapes incentives and stabilises cooperation. These ideas extend to multi-agent settings: Guéron et al. [12] established a folk theorem for repeated games with unequal discounting, which Chen and Takahashi [8] generalised to n-player games.

While these works focus on stateless repeated games, where each stage is identical and independent of past actions apart from discounting, our work introduces asymmetric discounting into stochastic games on graphs, where actions influence state transitions over time. In this setting, time preferences interact with dynamics and incentives, leading to fundamentally new forms of strategic complexity. To the best of our knowledge, such games have not been studied before.

The study of stable solution concepts in stochastic settings, including minimax values, Nash equilibria, and Stackelberg equilibria, has a long history [11, 22, 26]. Conitzer and Sandholm [9] formalised the problem of computing optimal strategies in leader-follower settings, but their model is restricted to stateless environments. Incentive equilibria, where the leader offers side payments to guide follower behaviour, were studied by Gupta et al. [15] in mean-payoff games. Their framework extends classical Stackelberg planning but does not involve discounting; strategies are evaluated using long-run average rewards, and players are assumed to have aligned temporal preferences.

Our work extends these ideas to dynamic settings by analysing ASGs on graphs. We show that, unlike in classical discounted games where stationary strategies suffice, optimal play in ASGs may require both memory and randomisation. This mirrors findings in other structured game models, such as pushdown games [7], where the recursive state space forces infinite-memory strategies, and finite-horizon MDPs [5], where achieving ε-optimality may require randomised or counting strategies.

2 Preliminaries

We write $f : A \to B$ for a function from A to B, and $f : A \rightharpoonup B$ for a *partial function*. A *probability distribution* over a finite set S is a function $d\colon S \to [0,1]$ such that $\sum_{s\in S} d(s) = 1$. Let $\mathcal{D}(S)$ denote the set of all probability distributions over S. A distribution $d \in \mathcal{D}(S)$ is called a *point distribution* if there exists some $s \in S$ such that $d(s) = 1$. For a distribution $d \in \mathcal{D}(S)$, we write $supp(d)$ for the *support* of d, defined as $supp(d) = \{s \in S \mid d(s) > 0\}$.

2.1 Stochastic Game Arena

▶ **Definition 2** (Stochastic Game Arena). *A* stochastic game arena *(SGA) $\mathcal{G}$ is a tuple $(S, A, T, R, s_0, S_{\mathsf{MAX}}, S_{\mathsf{MIN}})$, where:*

- *S is a finite set of states with a distinguished initial state $s_0 \in S$,*
- *A is a finite set of* actions*,*
- *$T\colon S \times A \rightharpoonup \mathcal{D}(S)$ is a* partial probabilistic transition function*,*
- *$R\colon S \times A \to \mathbb{Q}$ is a* reward function *representing the payoff from Player Min (or simply Min) to Player Max (or simply Max), with rewards encoded in binary.*
- *$\{S_{\mathsf{MAX}}, S_{\mathsf{MIN}}\}$ is a partition of S indicating states controlled by each player.*

For $s \in S$, let $A(s)$ denote the set of actions available at s. For $s, s' \in S$ and $a \in A(s)$, we write $p(s' \mid s, a)$ to denote $T(s, a)(s')$.

A *run* ρ of $\mathcal{G}$ is an ω-word $\langle s_0, a_1, s_1, \ldots\rangle \in S \times (A \times S)^\omega$ such that $p(s_{i+1}|s_i, a_{i+1}) > 0$ for all $i \geq 0$. A finite run is a finite such sequence, that is, a word in $S \times (A \times S)^*$. We write $\mathit{Runs}^{\mathcal{G}}$ (resp. $\mathit{FRuns}^{\mathcal{G}}$) for the set of runs (resp. finite runs) of the SGA $\mathcal{G}$ and $\mathit{Runs}^{\mathcal{G}}(s)$ (resp. $\mathit{FRuns}^{\mathcal{G}}(s)$) for the set of runs (resp. finite runs) of the SGA $\mathcal{G}$ starting from state s. We write $\mathit{last}(\rho)$ for the last state of a finite run ρ.

A stochastic game on an SGA $\mathcal{G}$ begins with a token placed in an *initial state* $s_0 \in S$. Players Max (he/him) and Min (she/her) construct an infinite run by taking turns to choose enabled actions, after which the token is moved to a successor state sampled from the selected distribution. A strategy for Max in $\mathcal{G}$ is a partial function $\sigma\colon \mathit{FRuns} \rightharpoonup \mathcal{D}(A)$, defined for $\rho \in \mathit{FRuns}$ if and only if $\mathit{last}(\rho) \in S_{\mathsf{MAX}}$, such that $\mathit{supp}(\sigma(\rho)) \subseteq A(\mathit{last}(\rho))$. A strategy τ for Min is defined analogously. We define the following classes of special strategies:

- A strategy μ is *pure* if $\mu(\rho)$ is a point distribution wherever it is defined; otherwise, μ is called a *mixed* strategy.
- A strategy μ is *stationary* if $\mathit{last}(\rho) = \mathit{last}(\rho')$ implies $\mu(\rho) = \mu(\rho')$ wherever μ is defined.
- A strategy is *positional* if it is both pure and stationary.
- A strategy μ is *counting* if, for any two runs ρ and ρ' of equal length, $\mathit{last}(\rho) = \mathit{last}(\rho')$ implies $\mu(\rho) = \mu(\rho')$ wherever μ is defined.

Let Σ_{MAX} and Σ_{MIN} be the sets of all strategies of Max and Min, respectively. Similarly, Π_{MAX} and Π_{MIN} denote the sets of all *positional* strategies of Max and Min, respectively. Let $\mathit{Runs}^{\mathcal{G}}_{\sigma,\tau}(s)$ denote the subset of runs $\mathit{Runs}^{\mathcal{G}}(s)$ starting from state s that are consistent with Max and Min following strategies σ and τ, respectively.

The behaviour of an SGA $\mathcal{G}$ under a strategy pair $(\sigma, \tau) \in \Sigma_{\mathsf{MAX}} \times \Sigma_{\mathsf{MIN}}$ is defined on a probability space $(\mathit{Runs}^{\mathcal{G}}_{\sigma,\tau}(s), \mathcal{F}_{\mathit{Runs}^{\mathcal{G}}_{\sigma,\tau}(s)}, \Pr^{\mathcal{G}}_{\sigma,\tau}(s))$ over the set of infinite runs $\mathit{Runs}^{\mathcal{G}}_{\sigma,\tau}(s)$ starting from state s. Regarding the underlying probability space: the sigma-algebra is initially defined over cylinder sets corresponding to finite runs and is extended to a full probability measure over infinite runs using the Ionescu-Tulcea extension theorem [18].

Given a random variable $f\colon \mathit{Runs}^{\mathcal{G}} \to \mathbb{R}$ over the infinite runs of $\mathcal{G}$, we denote by $\mathbb{E}^{\sigma,\tau}_s\{f\}$ the expectation of f w.r.t the probability space $(\mathit{Runs}^{\mathcal{G}}_{\sigma,\tau}(s), \mathcal{F}_{\mathit{Runs}^{\mathcal{G}}_{\sigma,\tau}(s)}, \Pr^{\mathcal{G}}_{\sigma,\tau}(s))$. For $n \geq 0$, we write X_n, Y_{n+1}, and $R_n = R(X_n, Y_{n+1})$ for the random variables corresponding to the n-th state, $(n{+}1)$-th action, and payoff to Max by Min at the n-th step.

2.2 Asymmetrically Discounted Stochastic Games

In many real-world scenarios, agents do not share identical time preferences. For example, in economic or multi-agent planning settings, one agent may be more far-sighted or patient than another because of structural incentives, cognitive traits, or mission-critical priorities. Classical models that assume symmetric discounting fail to capture such natural asymmetries. This motivates the study of games in which players use different discount factors, reflecting their individual perspectives on future rewards.

An asymmetrically discounted stochastic game $\mathcal{G}_{\alpha,\beta} = (\mathcal{G}, \alpha, \beta)$, with SGA $\mathcal{G} = (S, A, T, R, s_0, S_{\mathsf{MAX}}, S_{\mathsf{MIN}})$, is characterised by two rational discount factors $\alpha, \beta \in [0, 1)$. Without loss of generality, we assume that α is the larger discount factor. We refer to the player with discount factor α as Max, and to the other player as Min.

We say that Max's and Min's discount factors are *reasonably apart*, written $\alpha \gg \beta$, if $\frac{1}{\frac{\alpha}{\beta}-1}$ is polynomially bounded in the size of the input, or if $\beta = 0$. This condition can, for instance, be enforced by encoding the fractions in unary. Its purpose is only to exclude cases in which the two discount factors are barely distinguishable, such as $\alpha = \frac{k+1}{2k+1}$ and $\beta = \frac{k}{2k-1}$ when k is exponential in the size of the input.

When both players have the same discount factor, say α, we write $\mathcal{G}_\alpha$ for the resulting game. When clear from context, we write $\mathcal{G}$ for $\mathcal{G}_{\alpha,\beta}$.

The stochastic asymmetrically discounted payoff $\mathcal{D}^{\sigma,\tau}_{\mathsf{MAX}}(s)$ and $\mathcal{D}^{\sigma,\tau}_{\mathsf{MIN}}(s)$ for Max and Min, respectively, under the strategy pair $(\sigma, \tau) \in \Sigma_{\mathsf{MAX}} \times \Sigma_{\mathsf{MIN}}$ from state $s \in S$ are given as:

$$\mathcal{D}^{\sigma,\tau}_{\mathsf{MAX}}(s) = \mathbb{E}^{\sigma,\tau}_s \left\{ \lim_{N\to\infty} \sum_{i=0}^{N} \alpha^i R_i \right\} \quad \text{and} \quad \mathcal{D}^{\sigma,\tau}_{\mathsf{MIN}}(s) = \mathbb{E}^{\sigma,\tau}_s \left\{ \lim_{N\to\infty} \sum_{i=0}^{N} \beta^i R_i \right\}.$$

The objective of Max is to maximise the payoff $\mathcal{D}^{\sigma,\tau}_{\mathsf{MAX}}(s)$, while the objective of Min is to minimise $\mathcal{D}^{\sigma,\tau}_{\mathsf{MIN}}(s)$. As demonstrated in Section 1, asymmetrically discounted stochastic games (ASGs) are not determined; the upper and lower values may differ even in simple settings. For this reason, we explore alternative notions of strategic stability in such games.

2.3 Incentive Stackelberg Equilibria

We consider strategies for Max that, in addition to earning rewards, allow him to pay an *incentive* to Min. Such strategies are relevant in scenarios where Min optimises her own payoff rather than minimising Max's payoff. In this setting, Min deviates from a prescribed strategy only if doing so is profitable. By offering suitable incentives, Max can influence Min's decisions and achieve a higher payoff without reducing Min's payoff. To this end, Max employs an *incentive function*, which specifies the incentive paid to Min as a function of the history of play. Formally, an *incentive function* is a partial function $\iota\colon \mathit{FRuns} \rightharpoonup \mathbb{R}_{\geq 0}$. Let $\mathcal{I}$ denote the set of all incentive functions. An *incentivised strategy* for Max is a pair $(\iota, \sigma) \in \mathcal{I} \times \Sigma_{\mathsf{MAX}}$. Given a run $\rho \in \mathit{Runs}$, let $\iota_n(\rho)$ denote the incentive paid by Max at the nth step. At the ith step, Max pays $\iota_i(\rho)$ and Min receives $\iota_i(\rho)$, with the corresponding discounting applied. Equivalently, $-\iota_i(\rho)$ is added to Min's reward.[1] We write I_n for the random variable corresponding to the incentive at step n.

An incentivised strategy profile is a tuple $(\iota, \sigma, \tau) \in \mathcal{I} \times \Sigma_{\mathsf{MAX}} \times \Sigma_{\mathsf{MIN}}$. The payoff to both players under an incentivised strategy profile (ι, σ, τ) is given as:

$$\mathcal{D}^{\iota,\sigma,\tau}_{\mathsf{MAX}}(s) = \mathbb{E}^{\sigma,\tau}_s \left\{ \lim_{N\to\infty} \sum_{i=0}^{N} \alpha^i (R_i - I_i) \right\} \text{ and } \mathcal{D}^{\iota,\sigma,\tau}_{\mathsf{MIN}}(s) = \mathbb{E}^{\sigma,\tau}_s \left\{ \lim_{N\to\infty} \sum_{i=0}^{N} \beta^i (R_i - I_i) \right\}.$$

We consider *incentive Stackelberg equilibria* (ISE) as stable solutions of the game. In such an equilibrium, Min compares the payoff obtained by following the prescribed strategy and receiving the incentive with the payoff she can obtain by deviating without incentives. She deviates only if doing so yields a better payoff; otherwise, she follows the prescribed strategy. This can be interpreted as a contract: the incentive ensures that following the prescribed strategy is optimal for Min.

[1] Adding $-\iota_i(\rho)$ to Min's reward aligns with her objective to minimise.

When Max proposes an incentivised strategy profile (ι, σ, τ) to Min, she has two options: (i) play τ and obtain payoff $\mathcal{D}^{\iota,\sigma,\tau}_{\mathsf{MIN}}(s_0)$, or (ii) reject the proposal by playing some other strategy τ' and expect to obtain payoff $\sup_{\sigma' \in \Sigma_{\mathsf{MAX}}} \mathcal{D}^{\sigma',\tau'}_{\mathsf{MIN}}(s_0)$. Under this semantics, an incentivised strategy profile (ι, σ, τ) is an ISE, or *stable*, if Min does not gain by deviating. Recall that Min can only accept or reject the proposal wholesale. Thus,

$$\mathcal{D}^{\iota,\sigma,\tau}_{\mathsf{MIN}}(s_0) \leq \inf_{\tau' \in \Sigma_{\mathsf{MIN}}} \sup_{\sigma' \in \Sigma_{\mathsf{MAX}}} \mathcal{D}^{\sigma',\tau'}_{\mathsf{MIN}}(s_0).$$

An *optimal incentive Stackelberg equilibrium* (OIE) is an optimal incentivised profile that returns the best payoff to Max among all stable incentivised profiles. Formally, we define this optimal payoff $\mathsf{Val}_{\mathsf{OIE}}(s_0)$ from start vertex s_0 as:

$$\mathsf{Val}_{\mathsf{OIE}}(s_0) = \sup_{(\iota,\sigma,\tau)} \{\mathcal{D}^{\iota,\sigma,\tau}_{\mathsf{MAX}}(s_0) \,:\, (\iota, \sigma, \tau) \text{ is stable.}\}$$

This value is attained by at least one incentivised stable strategy profile. An OIE is a stable incentivised strategy profile $(\iota^*, \sigma^*, \tau^*)$ such that $\mathcal{D}^{\iota^*,\sigma^*,\tau^*}_{\mathsf{MAX}}(s_0) = \mathsf{Val}_{\mathsf{OIE}}(s_0)$. Given our semantics, where incentives are offered and accepted upfront and do not allow per-step negotiation, in fact it suffices to consider one-time incentive schemes.

▶ **Proposition 3** (One-off Incentives are Optimal.). *The value* $\mathsf{Val}_{\mathsf{OIE}}$ *can be obtained by an* OIE (ι, σ, τ)*, where* ι *is an incentive function such that Max pays a single incentive only at the very first step and zero in all following steps, i.e.,*

1. *for every run* ρ *we have* $\iota_i(\rho) = 0$ *for every step* $i > 0$ *and*
2. *for every pair of runs* $\rho, \rho' \in FRuns^{\mathcal{G}}(s)_0$*, we have* $\iota_0(\rho) = \iota_0(\rho')$*.*

Proof. Let (ι, σ, τ) be an OIE achieving $\mathsf{Val}_{\mathsf{OIE}}$. Define ι' by $\iota'_0(\rho) = \sum_{i=0}^{\infty} \beta^i \cdot \iota_i(\rho), \quad \iota'_j(\rho) = 0$ for all $j > 0$. Since Min's total discounted incentive remains unchanged, her incentive to follow τ is preserved. As τ is fixed, ι'_0 can be made constant across runs. Hence, (ι', σ, τ) is an OIE with a one-off, uniform incentive achieving $\mathsf{Val}_{\mathsf{OIE}}$. ◀

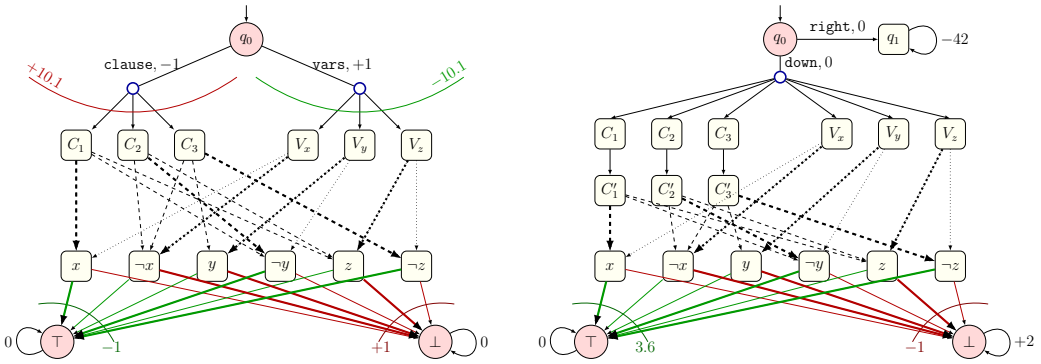
As a consequence, henceforth we will restrict our attention to incentivised strategies of this kind in the remaining of the paper. To simplify our terminology, with a slight abuse of notation we will use the term incentive to just denote ι, which is a number in $\mathbb{R}$. Hence, we define this optimal payoff from start vertex s_0 as follows:

$$\mathsf{Val}_{\mathsf{OIE}}(s_0) = \sup_{(\iota,\sigma,\tau)} \left\{ \mathcal{D}^{\sigma,\tau}_{\mathsf{MAX}}(s_0) - \iota \,:\, \mathcal{D}^{\sigma,\tau}_{\mathsf{MIN}}(s_0) - \iota \leq \inf_{\tau' \in \Sigma_{\mathsf{MIN}}} \sup_{\sigma' \in \Sigma_{\mathsf{MAX}}} \mathcal{D}^{\sigma',\tau'}_{\mathsf{MIN}}(s_0) \right\}.$$

Given the simplicity and appeal of stationary strategies, it is natural to ask whether stable equilibria can be found within this restricted class. In Section 3, we show that the problem becomes intractable. We consider two variants:

- **Optimal Stackelberg Equilibria.** Section 3 first examines the setting where Max is not permitted to offer incentives. In this case, a strategy profile $(\sigma, \tau) \in \Sigma_{\mathsf{MAX}} \times \Sigma_{\mathsf{MIN}}$ is an SE (or *stable*) if $\mathcal{D}^{\sigma,\tau}_{\mathsf{MIN}}(s_0) \leq \mathcal{D}^{\sigma,\tau'}_{\mathsf{MIN}}(s_0)$ for all $\tau' \in \Sigma_{\mathsf{MIN}}$. The corresponding decision problem, PosStackelbergIncentiveFree, asks whether there exists a stationary stable profile (σ, τ) such that $\mathcal{D}^{\sigma,\tau}_{\mathsf{MAX}}(s_0) \geq \lambda$ for a given threshold λ. We show this problem is intractable (Theorem 4) via a reduction from 3SAT.
- **Optimal Incentive Stackelberg Equilibria.** Section 3 then continues by strengthening this result by showing that the problem remains hard even when incentives are allowed. In the PosStackelbergIncentive problem, the goal is to determine whether there exists a stationary stable incentivised profile (ι, σ, τ) such that $\mathcal{D}^{\iota,\sigma,\tau}_{\mathsf{MAX}}(s_0) \geq \lambda$. The hardness proof builds upon the earlier reduction, adapted to incorporate incentive transfers (Theorem 6).

In Section 4, we present and analyse an algorithm for computing OIE under general strategies.



(a) Reduction for PosStackelbergIncentive-Free. **(b)** Reduction for PosStackelbergIncentive.

Figure 2 Reductions exemplified using the 3-CNF formula $\phi = (x \vee \neg y \vee z) \wedge (\neg x \vee \neg y \vee z) \wedge (\neg x \vee y \vee \neg z)$. Curved arcs indicate rewards associated with groups of edges. Edges without labels carry reward 0. Probabilistic nodes have outgoing transitions with uniform probability.

3 Intractability without Memory with or without Incentives

We start by considering the case where we disallow incentive and insist on positional strategies. To establish NP-hardness, we reduce from 3-SAT, exemplified in Figure 2(a). The reduction provides different pathways to literals: Min chooses to either go through a "clause" or a "vars" pathway, where a clause state or variable state, respectively, is selected uniformly at random. Max then chooses a literal from the clause or for the variable, respectively, and then goes to a $\top$ or $\bot$ sink. Fixing the discount factors to $\alpha = 0.9$ and $\beta = 0.1$, the payoffs are chosen such that a pathway through "clause" is always preferable to Max. Min, on the other hand, receives a payoff of at most 0 when going through "vars" and at least 0 when going through "clause", which means that going through "clause" requires a payoff of precisely 0 for both choices. This is only the case when *every* variable state selects a literal where Max moves to $\bot$, while every clause selects one of its literals from where Max moves to $\top$. For stationary strategies, this can be done if, and only if, the 3-CNF formula is satisfiable. The same holds if we allow counting strategies (that only depend on the number of steps taken), as the literals are visited after exactly two steps.

▶ **Theorem 4.** *Determining whether an ASG admits an optimal Stackelberg equilibrium with non-negative expected payoff for Max is NP-hard for both stationary and counting strategies, regardless of whether the strategies are pure or stochastic, when incentives are not allowed.*

For inclusion in NP, we can simply guess the stationary strategies when considering pure strategies. For stochastic strategies, we can encode their existence; however, we have no argument why the probabilities have to be succinctly representable (so that we could explicitly guess them), so that this only provides inclusion in the existential theory of the reals (ETR).

▶ **Theorem 5.** *The question whether ASGs admit a Stackelberg equilibrium with non-negative (or positive) payoff for Max with stationary policies is NP-complete for pure strategies and in ETR for stochastic strategies.*

Proof details are provided in Appendix B.

For stationary strategies, the problem remains hard even when incentives are allowed. The hardness proof uses games exemplified in Figure 2(b). Min always needs to be incentivised to choose "down" over "right", and upon choosing "down", a "clause" or "variable" state is chosen uniformly at random.

For the value of the paths, the payoff for Max and Min is higher when the play continues to $\top$, but it is chosen so that (for $\alpha = \frac{2}{3}$ and $\beta = \frac{1}{3}$) for the longer paths – which are the paths through the "clause" states – the advantage for Max outweighs the disadvantage for Min, while it is the other way round for the shorter paths – through the "variable" states. Consequently, the payout (after deducing the required incentive) is highest if Max can assure to turn to $\bot$ on the shorter paths (through the "variable" states) and to $\top$ through the longer paths (through the "clause" states). For stationary strategies, this can again be done if, and only if, the 3-CNF formula is satisfiable, e.g. by picking a satisfying assignment, turn to $\top$ from true and to $\bot$ from false literals, turning to the literal for each variable from the "variable" states, and to a true literal from each primed "clause" state.

▶ **Theorem 6.** *The problem of determining whether an ASG admits an optimal incentive Stackelberg equilibrium with expected payoff at least a given threshold for Max is NP-complete for pure stationary strategies, and in ETR and NP-hard for stochastic stationary strategies.*

Proof details are provided in Appendix C.

4 Optimal Incentive Stackelberg Equilibria

We develop an algorithm to compute an OIE in ASGs, whose complexity is no worse than the complexity of solving two-player zero-sum symmetrically discounted games. Let us fix an SGA $\mathcal{G} = (S, A, T, R, s_0, S_{\mathsf{MAX}}, S_{\mathsf{MIN}})$ with discount factors α and β for Max and Min, respectively. We give a description of the algorithm in several steps. A summary of our algorithm for computing OIE is provided at Algorithm 1.

4.1 Step 1: Compute Threshold Value for Min

The very first step of our algorithm is to determine the best payoff Min can obtain if she decides to reject the incentivised strategy profile proposed by Max. From the definition of OIE, in this case, Min treats Max as an adversary, and plays the optimal strategy assuming the worst case behaviour from Max. Essentially this is equivalent to computing the optimal strategy of Min in the classic two-player symmetrically discounted game with her discount factor β. We solve the game $\mathcal{G}_\beta$ and get a positional optimal strategy τ_β of Min, along with the value $\overline{V}_{\mathsf{MIN}} = \inf_{\tau \in \Sigma_{\mathsf{MIN}}} \sup_{\sigma \in \Sigma_{\mathsf{MAX}}} \mathbb{E}_{s_0}^{\sigma,\tau} \left\{ \lim_{N \to \infty} \sum_{i=0}^{N} \beta^i R_i \right\}$. If Min rejects Max's offer, Max can play adversarially and ensure that Min receives at most $\overline{V}_{\mathsf{MIN}}$ by playing his optimal strategy in $\mathcal{G}_\beta$. Consequently, in any OIE, Min's payoff is at most $\overline{V}_{\mathsf{MIN}}$.

▶ **Observation 7.** *For any stable incentivised profile (ι, σ, τ), we have that $\mathcal{D}_{\mathsf{MIN}}^{\iota,\sigma,\tau}(s_0) \leq \overline{V}_{\mathsf{MIN}}$.*

The subsequent steps of our algorithm focus on the case when Min decides to follow the incentivised strategy profile proposed by Max. In this case, we naturally assume that players Max and Min play cooperatively, ensuring that Max maximises his payoff while minimising the incentive, under the constraint that the payoff of Min does not exceed $\overline{V}_{\mathsf{MIN}}$.

4.2 Step 2: Optimal Long Term Cooperation Value for Max

We compute the long-term behaviour of an OIE when Min agrees to follow the strategy profile proposed by Max. This is a positional pure strategy profile that both players eventually follow in an OIE. To this end, we show later that there is an initial finite horizon – a bound $\kappa \in \mathbb{N}$, so that after κ steps, it is optimal for both Max and Min to play this profile. Determining such an endgame profile is done in two steps.

We first compute the best case scenario for Max with cooperation of Min. For this, we consider the one-player maximiser game over arena $\mathcal{G}' = (S, A, T, R, s_0, S, \emptyset)$ with discount factor α, where all the vertices in $\mathcal{G}$ are now controlled by Max. Essentially, in this game Max makes choices for Min as well. We solve the one-player discounted payoff game $\mathcal{G}'_\alpha$ and obtain the value vector V_α^{MAX}, along with the set of all optimal actions (maximally-permissive set) $E \subseteq S \times A$. Next, we form the game arena $\mathcal{G}'' = (S, A, T_E, R, s_0, \emptyset, S)$, which is the arena where only optimal actions are enabled and every state is controlled by Min. Since any strategy profile in $\mathcal{G}'_\alpha$ fetches the optimal payoff to Max, our next step is to find strategy profiles among these, that are optimal for Min. For this, we consider the one-player minimiser game $\mathcal{G}''_\beta$. Solving this game, we obtain the value vector V_β^{MIN} along with the set of all (positional) optimal strategies of Min. Let $\hat{\Sigma} \subseteq \Pi_{\mathsf{MAX}} \times \Pi_{\mathsf{MIN}}$ be the set of all positional strategy profiles that are optimal for Max and Min in games $\mathcal{G}'_\alpha$ and $\mathcal{G}''_\beta$ respectively.

▶ **Observation 8.** *In the game $\mathcal{G}_{\alpha,\beta}$, for any strategy profile $(\sigma, \tau) \in \hat{\Sigma}$ and any $s \in S$, $\mathcal{D}^{\sigma,\tau}_{\mathsf{MAX}}(s) = V_\alpha^{\mathsf{MAX}}(s)$ and $\mathcal{D}^{\sigma,\tau}_{\mathsf{MIN}}(s) = V_\beta^{\mathsf{MIN}}(s)$.*

At the start node s_0, if $V_\beta^{\mathsf{MIN}}(s_0) \leq \overline{V}_{\mathsf{MIN}}$, then any strategy profile (σ, τ) from $\hat{\Sigma}$, which gives the values $V_\alpha^{\mathsf{MAX}}(s_0)$ and $V_\beta^{\mathsf{MIN}}(s_0)$ to Max and Min respectively is an ISE with incentive $\iota = 0$. This is the best possible payoff Max can obtain in this game and the arena induced by optimal strategies of Max also favours Min to improve her payoff compared to her worst case payoff on deviation.

▶ **Observation 9.** *If $V_\beta^{\mathsf{MIN}}(s_0) \leq \overline{V}_{\mathsf{MIN}}$ then for any $(\sigma^*, \tau^*) \in \hat{\Sigma}$, $(0, \sigma^*, \tau^*)$ is an* OIE. *In this case, (σ^*, τ^*) is an* SE.

Hence, in this case, our algorithm terminates and outputs $(0, \sigma^*, \tau^*)$ as an OIE. On the other hand, if $V_\beta^{\mathsf{MIN}}(s_0) > \overline{V}_{\mathsf{MIN}}$, then $(0, \sigma, \tau)$ for any $(\sigma, \tau) \in \hat{\Sigma}$ is not an OIE (is not either an ISE, and (σ, τ) not a SE), since Min can improve her outcome playing τ_β. Hence, Max would need to incentivise Min to play a strategy that matches her deviation payoff and also is better for Max.

Denote $I_{\texttt{ceil}} = V_\beta^{\mathsf{MIN}}(s_0) - \overline{V}_{\mathsf{MIN}}$. Observe that, $I_{\texttt{ceil}}$ is the maximum incentive Max would have to possibly pay to Min, in order to convince Min to not play τ_β. But this is not necessarily optimal for Max. In other words the profile $(I_{\texttt{ceil}}, \sigma, \tau)$ for any $(\sigma, \tau) \in \hat{\Sigma}$ is an ISE, but not necessarily an OIE. In such strategy profiles, Max gets a payoff $V_\alpha^{\mathsf{MAX}}(s_0) - I_{\texttt{ceil}}$ and Min gets $V_\beta^{\mathsf{MIN}}(s_0) - I_{\texttt{ceil}} = \overline{V}_{\mathsf{MIN}}$. In order to obtain stable profiles with better payoff to Max, we will see that Max and Min can play differently from profiles in $\hat{\Sigma}$ for few steps in some initial horizon, following which they can no longer improve and resort to any profile in $\hat{\Sigma}$. Next, we will see that such a horizon κ exists and subsequently compute it.

4.3 Step 3: Computing the Beneficial Deviation Horizon

Since we still assume that both players are cooperating, once again Max chooses actions for both players. At each node $s \in S$ and for every action $a \in A(s)$, Max evaluates the decision to choose an action a instead of the action prescribed by some profile in $\hat{\Sigma}$ i.e., a

deviation in the very first step may yield improved outcomes for Min, while Max might lose; but overall this can result in Max paying less incentive to Min. To quantify the impact of such a deviation Max uses two quantities: he evaluates the change in his one-step payoff on choosing a, using the disadvantage function [3, 27] $\mathsf{MaxDisAdv}(s,a)$:

$$\mathsf{MaxDisAdv}(s,a) = V_\alpha^{\mathsf{MAX}}(s) - \big(R(s,a) + \alpha \sum_{s' \in S} p(s'|s,a) \cdot V_\alpha^{\mathsf{MAX}}(s')\big).$$

Note that $\mathsf{MaxDisAdv}(s,a)$ is 0 when a conforms to some strategy profile in $\hat{\Sigma}$ and positive otherwise. Second, Max evaluates the change in Min's payoff, or equivalently, the amount saved from the incentive he must pay, using the advantage function [3, 27] $\mathsf{MinAdv}(s,a)$:

$$\mathsf{MinAdv}(s,a) = V_\beta^{\mathsf{MIN}}(s) - \big(R(s,a) + \beta \sum_{s' \in S} p(s'|s,a) \cdot V_\beta^{\mathsf{MIN}}(s')\big).$$

Note that $\mathsf{MinAdv}(s,a)$ is 0 for actions conforming to strategy profiles in $\hat{\Sigma}$, negative in other actions from game $\mathcal{G}''$, and can be anything elsewhere. When Max plays action a at state s, in the k^{th} step, its discounted one-step benefit of choosing a at k step is given by:

$$\mathsf{benefit}_k(s,a) = \beta^k \mathsf{MinAdv}(s,a) - \alpha^k \mathsf{MaxDisAdv}(s,a).$$

We will later show that for reasonably apart discount factors there is a small $\kappa \in \mathcal{N}$ s.t., for all $j \geq \kappa$ and $\forall s \in S, a \in A(s)$, $\mathsf{benefit}_j(s,a) \leq 0$. Once this is the case, for computing OIE we can focus only on strategies that do not deviate from $\hat{\Sigma}$ after κ steps. With this in mind, we consider *k-eventually positional strategy profiles*, which are strategies where Max and Min, from the kth step onward play according to some profile in $\hat{\Sigma}$, and anything before that. When Max and Min play a k-eventually positional strategy profile, let A_n denote the advantage to Min at step n and B_n denote the discounted benefit at step n.

▶ **Lemma 10.** *Let (σ_k, τ_k) be a k-eventually positional strategy profile of Max and Min. Then $(I_{\mathtt{ceil}} - \mathbb{E}_{s_0}^{\sigma_k,\tau_k}(\sum_{i=0}^{k-1} \beta^i A_i), \sigma_k, \tau_k)$ is a stable incentivised strategy profile where the payoff to Max is $(V_\alpha^{\mathsf{MAX}}(s_0) - I_{\mathtt{ceil}}) + \mathbb{E}_{s_0}^{\sigma_k,\tau_k}(\sum_{i=0}^{k-1} B_i)$.*

From Lemma 10 it follows that as long as the k-th step expected benefit is positive, it is beneficial for Max. But since $\beta < \alpha$, and there are only finitely many $\mathsf{MaxDisAdv}$ and MinAdv values, the benefit becomes negative after some step κ. From linearity of expectation, the payoff does not increase further and the optimal strategy profile would be κ-eventually positional. Now to estimate κ, let $q = \max_{(s,a) \in S \times A} \frac{\mathsf{MinAdv}(s,a)}{\mathsf{MaxDisAdv}(s,a)}$ (Excluding (s,a) if $\mathsf{MaxDisAdv}(s,a) = 0$). Note that q is exponentially bounded by value, and polynomially by length, in the input size.

▶ **Observation 11.** *If, for some k, we have $\frac{\beta^k q}{\alpha^k} \leq 1$, then for every $j \geq k$, every state $s \in S$, and every action $a \in A(s)$, we have $\mathsf{benefit}_j(s,a) \leq 0$.*

It is easy to see that $\kappa = \lceil \log_{\alpha/\beta} q \rceil = \left\lceil \frac{\ln q}{\ln(\alpha/\beta)} \right\rceil$. Moreover, since we have assumed $\alpha \gg \beta$, κ is polynomial in the input size. Thus there is a κ-eventually positional OIE, which we compute next.

4.4 Step 4: Synthesising the Optimal Initial Strategy via LP

In the final step, we write a linear program (LP) called SynthInitialOptimal to compute a strategy profile that is optimal for the first κ steps. Starting from start vertex s_0, we unfold the game arena for κ steps and compute the expected benefit and savings only for this horizon. The reason that we can unfold the arena, is the crucial observation that in the first κ steps, counting is enough.

▶ **Observation 12.** *There is a κ-eventually positional* OIE*, which has counting strategy in the first κ step.*

Hence, in the LP SynthInitialOptimal, for all $s \in S, a \in A(s)$, we have a variable $x^i_{s,a}$ that denotes the probability that in the ith step, the action a is played from state s. At Equation (4) only the reachable $x^i_{s,a}$ are considered.

$$\text{maximise} \sum_{i=0}^{\kappa-1} \sum_{s\in S} \sum_{a\in A(s)} x^i_{s,a}\mathsf{benefit}_i(s,a) \text{ subject to:} \tag{1}$$

$$0 \le x^i_{s,a} \le 1 \qquad \forall i, \forall s, \forall a \in A(s) \tag{2}$$

$$\sum_{a\in A(s_0)} x^0_{s_0,a} = 1 \tag{3}$$

$$\sum_{a\in A(s)} x^{i+1}_{s,a} = \sum_{s'\in S} \sum_{a'\in A(s')} p(s|s',a')x^i_{s',a'} \qquad \forall s \in S, \forall\, 0 \le i < \kappa \tag{4}$$

$$I_{\mathtt{ceil}} - \sum_{i=0}^{\kappa-1} \sum_{s\in S} \sum_{a\in A(s)} x^i_{s,a}\beta^i\mathsf{MinAdv}(s,a) \ge 0 \tag{5}$$

▶ **Proposition 13.** *The optimal solution θ^* to SynthInitialOptimal, gives a κ-eventually positional incentivised strategy profile $(\iota^*, \sigma_{\theta^*}, \tau_{\theta^*})$, which is an* OIE*.*

Proof. A unique strategy can be extracted from any valuation satisfying Equations (2)–(4). The L.H.S. in Equation (5) is the incentive, ensuring that the incentive is non-negative. Finally, by Lemma 10, Equation (1) suffices for finding an optimal profile. ◀

Complexity. The overall complexity of computing an OIE breaks down into three main components. First, the initial phase of the game can be unravelled into a directed acyclic graph, enabling the construction of a LP of polynomial size. Second, this LP determines the early-game behaviour, while the long-run behaviour converges to a stationary profile by solving one-player games. The only non-polynomial step involves solving a classical two-player discounted game with Player Min's (smaller) discount factor, which determines her fallback guarantee and thus the incentive threshold. Since solving discounted games becomes easier with bounded discount factors this step remains practical in most settings [16]. As for the overall complexity, classical two-player discounted game can be solved in UP ∩ co-UP [17] and in the following steps of the algorithm there is no non-determinism involved.

▶ **Theorem 14.** *Given a ASG $\mathcal{G}_{\alpha,\beta}$ where α and β are reasonably apart, an Optimal Incentive Stackelberg Equilibrium in this game can be computed in UP ∩ co-UP.*

Algorithm 1 Compute Optimal Incentive Stackelberg Equilibria.

1: **Input:** $\mathcal{G}, \alpha, \beta$
2: **Output:** optimal incentive Stackelberg equilibrium $(\iota^*, \sigma^*, \tau^*)$
3: $\overline{V}_{\mathsf{MIN}} \leftarrow$ value of $\mathcal{G}_\beta$
4: $\mathcal{G}' \leftarrow \mathcal{G}$ with all vertices controlled by Max
5: Solve $\mathcal{G}'_\alpha$ and obtain V_α^{MAX}
6: $\mathcal{G}'' \leftarrow \mathcal{G}'$ restricted to Max-optimal edges in $\mathcal{G}'_\alpha$, with all vertices controlled by Min
7: Solve $\mathcal{G}''_\beta$ and obtain V_β^{MIN}
8: $\hat{\Sigma} \leftarrow$ optimal Max–Min profiles induced by $\mathcal{G}'_\alpha$ and $\mathcal{G}''_\beta$
9: **if** $V_\beta^{\mathsf{MIN}}(s_0) \le \overline{V}_{\mathsf{MIN}}$ **then**
10: $\quad(\sigma^*, \tau^*) \leftarrow$ any profile in $\hat{\Sigma}$
11: $\quad$**return** $(0, \sigma^*, \tau^*)$
12: **end if**
13: Compute $\mathsf{MaxDisAdv}(s,a)$ and $\mathsf{MinAdv}(s,a)$ for all $s \in S$ and $a \in A(s)$
14: $q \leftarrow \max\left\{\frac{\mathsf{MinAdv}(s,a)}{\mathsf{MaxDisAdv}(s,a)} \;:\; s \in S,\ a \in A(s),\ \mathsf{MaxDisAdv}(s,a) > 0\right\}$
15: $\kappa \leftarrow \min\left\{k \mid \frac{\beta^k q}{\alpha^k} \le 1\right\}$
16: Solve SynthInitialOptimal on the first κ-step unfolding of $\mathcal{G}$ and extract $(\iota^*, \sigma^*, \tau^*)$
17: **return** $(\iota^*, \sigma^*, \tau^*)$

5 Experimental Results

In this section, we report experimental results from our implementation of Section 4 available at [2]. All experiments were run in Python 3.12.3 on a Windows 11 machine with an AMD Ryzen 5 Pro 7535U CPU at 2.90 GHz and 15.6 GB RAM.

We evaluate our implementation on random games and on three PRISM-games benchmarks [19, 6], adapted to our model with suitable rewards: Avoid-the-Observer, Dice, and Hallway–Human. The Dice benchmark is played over a bounded number of rounds, whereas Avoid-the-Observer and Hallway–Human are grid-based games. We vary the grid size, the number of rounds, and the pair of discount factors. The results are reported in Table 1. The timeout was set to 10 minutes, and "–" denotes timeout. For each model, we report the number of states in the corresponding ASG, the discount factors, the runtime, the counting horizon κ, and the incentive paid by Max to Min in the OIE.

- The Avoid-the-Observer (AV) game [6] models a pursuit–evasion scenario on an $N \times N$ grid, where the intruder aims to avoid capture while reaching a fixed exit or collecting a treasure. We add step costs and rewards to capture this objective, and run the benchmark on grids of sizes 10, 15, and 18.
- The Dice game [19] is played between two players, each of whom may roll a die for at most N rounds. The winner is the player with the higher total outcome. We add rewards to penalise prolonged play and encode the final comparison of outcomes, and use instances with 10, 20, and 50 rounds. For these games, Algorithm 1 terminates at Line 11, so Max need not offer any incentive.
- The Hallway–Human (HW) game [6] models a robot navigating an $N \times N$ grid subject to probabilistic motion failures and damage, while interacting with a human whose behaviour alternates between stochastic and adversarial choices. The robot's objective is to reach the human while minimising the probability of damage. We run this benchmark on grids of sizes 5, 8, and 10. When the discount factors are close, namely $\alpha = 0.55$ and $\beta = 0.50$, the counting horizon becomes large; for the 8×8 and 10×10 grids, this increases the LP size enough to cause timeout.

Table 1 Results from running Section 4 on three benchmarks.

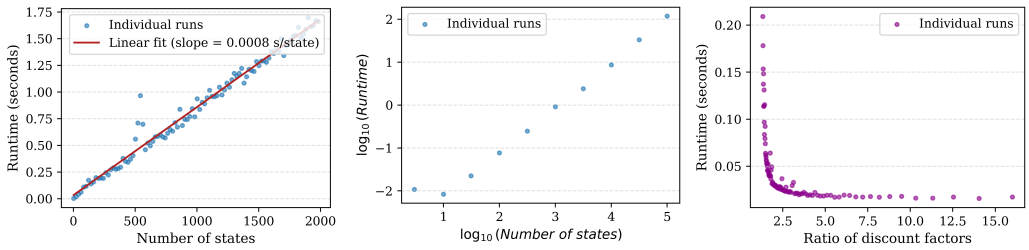
Model	ASG info			Runtime (sec)	κ	Incentive	Player's Payoff	
	#states	α	β				Max	Min
AV[10 × 10]	19803	0.9	0.1	38.86	2	0.05	47.36	9.04
		0.67	0.33	14.64	3	0.59	16.33	9.51
		0.55	0.50	15.71	17	1.52	12.90	10.47
AV[15 × 15]	100803	0.9	0.1	238.08	2	10.05	47.36	-0.95
		0.67	0.33	103.83	3	0.58	16.3	-0.48
		0.55	0.50	111.61	17	11.52	12.90	0.47
AV[18 × 18]	209307	0.9	0.1	515.59	1	10.01	47.36	-0.90
		0.67	0.33	222.14	2	10.02	16.33	0.07
		0.55	0.50	237.44	23	10.15	12.90	1.84
Dice[10]	2752	0.9	0.1	0.43	0	0	-9.05	-5.04
		0.67	0.33	0.36	0	0	-7.24	-5.54
		0.55	0.50	0.49	0	0	-6.51	-6.25
Dice[20]	9682	0.9	0.1	1.49	0	0	-9.05	-5.04
		0.67	0.33	1.64	0	0	-7.24	-5.54
		0.55	0.50	2.02	0	0	-6.51	-6.25
Dice[50]	55672	0.9	0.1	9.92	0	0	-9.05	-5.04
		0.67	0.33	11.65	0	0	-7.24	-5.54
		0.55	0.50	13.78	0	0	-6.51	-6.25
HW[5 × 5]	7602	0.9	0.1	23.25	0	0	3.76	-0.01
		0.67	0.33	10.80	2	0.00012	0.06	-0.01
		0.55	0.50	17.02	19	0.0048	-0.01	-0.02
HW[8 × 8]	49410	0.9	0.1	206.25	0	0	1.76	-0.01
		0.67	0.33	84.08	2	0	-0.02	-0.01
		0.55	0.50	–	35	–	-	–
HW[10 × 10]	120402	0.9	0.1	501.4	0	0	1.04	-0.01
		0.67	0.33	207.60	0	0	-0.03	-0.01
		0.55	0.50	–	35	–	–	–

Across the three benchmarks, the LP size scales with the number of states multiplied by $\kappa + 1$. Some instances therefore produce LPs with millions of variables, yet the running times remain manageable in most cases.

Scalability on Random Games. For random games, we generate balanced bipartite games with K states for each player. We vary K over a moderate range, up to 1000, in Figure 3 (left), over a wider range from 3 to 10^5 in Figure 3 (middle), and fix the size at 30 while varying the ratio between α and β in Figure 3 (right). The random-game experiments show that runtime grows linearly with game size. Although the theoretical bottleneck is computing the minimum payoff Min can obtain, this step is fast in practice because β is the smaller discount factor. The observed trend is therefore best understood as linear growth in the generated LP size, which modern LP solvers handle efficiently in these instances. Finally, as the ratio between α and β approaches 1, the counting horizon κ increases, explaining the runtime rise in Figure 3 (right).

6 Conclusion

We introduced the concept of asymmetric discounting to two-player stochastic games. This notion arises naturally in many settings and has significant implications for the structure and analysis of discounted games. In particular, asymmetry in time preferences breaks the standard symmetry between players – effectively granting one player a longer planning horizon. This positions the more patient player as a natural leader, making variants of Stackelberg equilibria the most suitable solution concepts in such contexts.



(a) Moderate sizes. **(b)** Wide size range. **(c)** Discount separation.

Figure 3 Running time grows linearly with the number of states for ASGs up to $2K$ states and for sizes from 3 to 10^5; it decreases as the discount-factor separation grows.

Our focus has been on *optimal incentive Stackelberg equilibria*, where the leader can offer direct utility transfers to the follower to influence their strategic choices. This added expressive power enables efficient handling of heterogeneous discounting: it yields rational and computationally tractable equilibria, with potential deviations from optimal play confined to a short initial prefix. Except for solving a classical discounted payoff game to determine the follower's baseline value, the computation remains efficient.

Asymmetric discounting also introduces several non-classical phenomena. Optimal strategies may require both memory and randomisation, and restricting strategies to be positional significantly increases computational complexity. Moreover, heterogeneous discounting allows both players to find infinite paths profitable, further complicating equilibrium reasoning. These findings open up new avenues for research, including broader classes of Nash and Stackelberg equilibria, and variants of incentive equilibria where compliance may be enforced on an action-by-action basis rather than globally.

References

1 M. Agranov, J. Kim, and L. Yariv. Coordination with differential time preferences: Experimental evidence. *American Economic Review: Insights*, pages 543–57, December 2024.

2 S. Bahmani, S. Paul, S. Schewe, S. Tasdighi Kalat, and A. Trivedi. Experiments for asymmetrically-discounted stochastic games, June 2026. doi:10.5281/zenodo.21101820.

3 L. C. Baird. Advantage updating, 1993.

4 Moshe Bar. Wait for the second marshmallow? future-oriented thinking and delayed reward discounting in the brain. *Neuron*, pages 4–5, 2010.

5 K. Chatterjee and R. Ibsen-Jensen. Strategy complexity of finite-horizon Markov decision processes and simple stochastic games. In *Proc. of MEMICS*, pages 106–117, 2012.

6 K. Chatterjee, J.-P. Katoen, M. Weininger, and T. Winkler. Stochastic games with lexicographic reachability-safety objectives. In *Proc. of CAV*, LNCS, pages 398–420, 2020.

7 K. Chatterjee and Y. Velner. Mean-payoff pushdown games. In *Proc. of LICS*, 2012.

8 B. Chen and S. Takahashi. A folk theorem for repeated games with unequal discounting. *Games and Economic Behavior*, pages 571–581, 2012.

9 V. Conitzer and T. Sandholm. Computing the optimal strategy to commit to. In *Proc. of the 7th ACM conference on Electronic commerce*, pages 82–90, 2006.

10 A. Dasgupta and S. Ghosh. Self-accessibility and repeated games with asymmetric discounting. *Journal of Economic Theory*, page 105312, 2022.

11 Jerzy A Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 1997.

12 Y. Guéron, T. Lamadon, and C. D Thomas. On the folk theorem with one-dimensional payoffs and different discount factors. *Games and Economic Behavior*, pages 287–295, 2011.

13 A. Gupta and S. Schewe. It pays to pay in bi-matrix games-a rational explanation for bribery. *Proc. of IFAAMAS*, 2015.

14 A. Gupta and S. Schewe. Buying optimal payoffs in bi-matrix games. *Games*, page 40, 2018. doi:10.3390/G9030040.

15 A. Gupta, S. Schewe, A. Trivedi, M. S. K. Deepak, and B. K. Padarthi. Incentive Stackelberg mean-payoff games. In *Proc. of SEFM*, pages 304–320, 2016.

16 T. D. Hansen, P. B. Miltersen, and U. Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *J. ACM*, pages 1:1–1:16, 2013. doi:10.1145/2432622.2432623.

17 Marcin Jurdzinski. Deciding the winner in parity games is in $UP \cap co$-UP. *Inf. Process. Lett.*, pages 119–124, 1998.

18 Olav Kallenberg. *Foundations of modern probability*. Springer, 1997.

19 M. Kwiatkowska, G. Norman, D. Parker, and C. Wiltsche. Prism-games 3.0: Stochastic game verification with concurrency, equilibria and time. In *Proc. of CAV*, pages 485–497, 2020.

20 E. Lehrer and A. Pauzner. Repeated games with differential time preferences. *Econometrica*, pages 393–412, 1999.

21 Fink A M. Equilibrium in a stochastic n-person game. *Hiroshima Math J*, 28:89–93, 1964.

22 J.-F. Mertens and A. Neyman. Stochastic games. *Int. J. Game Theory*, pages 53–66, 1981.

23 Walter Mischel. *The marshmallow test: Understanding self-control and how to master it*. Random House, 2014.

24 M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

25 A. Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, pages 97–109, 1982.

26 L. S. Shapley. Stochastic games. *Proc. Nat. Acad. Sci. U.S.A.*, pages 1095–1100, 1953.

27 R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

28 S. Tasdighi Kalat, S. Sankaranarayanan, and A. Trivedi. What is your discount factor? In *International Conference on Quantitative Evaluation of Systems and Formal Modeling and Analysis of Timed Systems*, pages 322–336, 2024.

29 Heinrich von Stackelberg. *Marktform und Gleichgewicht*. Springer, Vienna, 1934.

A Selected Worked-out Examples

Details on game in Figure 1a with $\alpha = 0.9$ and $\beta = 0.1$

Non-determinacy. The optimal value to Min when she plays first is: $\inf_\tau \sup_\sigma \mathcal{D}^{\sigma,\tau}_{\mathsf{MIN}}(q_0) = 0$. And this value is given by strategy τ_1 where Min plays b at q_0. The optimal value to Max when Min plays b at q_0 is: $V_0 = \sup_\sigma \mathcal{D}^{\sigma,\tau_1}_{\mathsf{MAX}}(q_0) = 0$. Now when Max plays first, consider the strategy where at q_1 Max plays a with probability p and b with probability $1-p$. The payoff to both Max and Min when Min chooses b is 0. But if Min chooses a, then the payoff to Min is $\frac{10-19p}{90-p}$, and the payoff to Max is $\frac{9-9.9p}{1-0.81p}$. The payoff to Min is 0 when $p = \frac{10}{19}$, hence for $p = \frac{10}{19}$, Min has no incentive to deviate from a to b. Moreover, the payoff to Max under this strategy is $\frac{720}{109}$. If $p < \frac{10}{19}$, then Min prefers to deviate to b, and Max's payoff is 0. Hence Max can choose $p > \frac{10}{19}$ to incentivise Min from not deviating. Since Max's payoff is decreasing in p, the optimal payoff to Max when playing first is $V_1 = \frac{720}{109}$. Therefore, the values V_0 and V_1 are not equal.

Value with incentive for Max. When Min chooses a at q_0 and Max chooses b at q_1, the payoff for Min is $\frac{0.1}{1-0.9} = \frac{1}{9}$. Hence for an incentive $\frac{1}{9} + \epsilon$ for some $\epsilon \geq 0$, Max can incentivise Min to choose a at q_0, since payoff to Min is now $-\epsilon < 0$. In this scenario, the payoff to Max is $\frac{0.9}{1-0.9} - (\frac{1}{9} + \epsilon) = (\frac{80}{9} - \epsilon)$. For an appropriate ϵ, this is greater than optimal payoff to Max without incentive which is 6.6. In fact $\epsilon = 0$ is enough, since this gets Min 0, which is no worse than what she would get by not following the proposed profile.

B Complexity of Stackelberg Equilibria with Positional or Counting Strategies without Incentives

In this appendix we study the setting where Player Max is not permitted to offer incentives. Recall that in this case a strategy profile $(\sigma, \tau) \in \Sigma_{\mathsf{MAX}} \times \Sigma_{\mathsf{MIN}}$ is *stable* if $\mathcal{D}^{\sigma,\tau}_{\mathsf{MIN}}(s_0) \leq \mathcal{D}^{\sigma,\tau'}_{\mathsf{MIN}}(s_0)$ for all $\tau' \in \Sigma_{\mathsf{MIN}}$. The corresponding decision problem, PosStackelbergIncentiveFree, asks whether there exists a stationary stable profile (σ, τ) such that $\mathcal{D}^{\sigma,\tau}_{\mathsf{MAX}}(s_0) \geq \lambda$ for a given threshold λ. We show that PosStackelbergIncentiveFree is NP-complete when strategies are restricted to be pure stationary and NP-hard when restricted to counting strategies. To establish hardness, we reduce from 3-SAT.

Proof. To prove Theorem 4, we construct a simple ASG from a 3-SAT problem, where the first choice is made by Min and all other choices are given to Max. An example is provided in Figure 2a. The payoffs are chosen such that Min's valuation of all runs that turn left are non-negative, while those of runs that turn right are non-positive, so that she would only agree to turn left if the value of all paths is 0 and this can be done if, and only if, we encode a satisfying assignment.

Construction. For a CNF formula ϕ over a set $X = \{x_1, x_2, \ldots, x_n\}$ of n variables that uses the set of m clauses $C = \{C_1, C_2, \ldots, C_m\}$. Each clause $C_i \in C$ contains three literals, denoted C_i^1, C_i^2, C_i^3. We construct an ASG $G_\phi = (S, A, T, R, q_0, S_{\mathsf{MAX}}, S_{\mathsf{MIN}})$ from ϕ where:

- $S = \{q_0, \top, \bot\} \cup C \cup \{V_x : x \in X\} \cup \{x, \neg x : x \in X\}$ is the set of states;
- $A = \{\texttt{clause}, \texttt{vars}\} \cup \{1, 2, 3\} \cup \{v, \neg v\} \cup \{\top, \bot\} \cup \{\star\}$;
- The transition function T is defined as:
  - $T(q_0, \texttt{clause})(s') = \frac{1}{m}$ if $s' \in C$ and is 0 otherwise;
  - $T(q_0, \texttt{vars})(s') = \frac{1}{n}$ if $s' \in V_x$ and is 0 otherwise;
  - $T(s, i)(\ell) = 1$ if $s \in C$, $i \in \{1, 2, 3\}$ and $\ell = s^i$;
  - $T(s, v)(s') = 1$ if $s = V_x$, $s' = x$ and $T(s, \neg v)(s') = 1$ if $s = V_x$, $s' = \neg x$;
  - $T(s, \top, \top) = 1$ $T(s, \bot, \bot) = 1$ and if $s \in \{x, \neg x : x \in X\}$;
  - $T(\top, \star, \top) = 1$ and $T(\bot, \star, \bot) = 1$ for all $a \in A$;
- $S_{\mathsf{MIN}} = \{q_0\}$ and $S_{\mathsf{MAX}} = S \setminus S_{\mathsf{MIN}}$.
- $R(s, \texttt{clause}) = -1$, $R(s, \texttt{vars}) = 1$, $R(s, \texttt{i}) = 10.1$ where $i \in \{1, 2, 3\}$, $R(s, v) = R(s, \neg v) = -10.1$, $R(s, \top) = -1$, $R(s, \bot) = +1$, and $R(s, \star) = 0$.

The discount factors of Max and Min be $\alpha = 0.9$ and $\beta = 0.1$, respectively.

Example. Consider the formula $\phi = (x \vee \neg y \vee z) \wedge (\neg x \vee \neg y \vee z) \wedge (\neg x \vee y \vee \neg z)$. Here for example $C_1 = (x \vee \neg y \vee z)$ and $C_1^1 = x$ and $C_1^2 = \neg y$. The corresponding game arena G_ϕ based on ϕ is shown in Figure 2a.

To understand the idea behind the construction, notice that if Min chooses variable nodes (right side), then all possible runs have the following valuations:

- Min's payoff: $(+1) + (-10.1\beta) + (\pm\beta^2) = -0.02$ or 0; and
- Max's payoff: $(+1) + (-10.1\alpha) + (\pm\alpha^2) = -7.28$ or -8.90.

If Min chooses clause nodes (left side) then all possible runs have the following valuations:

- Min's payoff: $(-1) + (10.1\beta) + (\pm\beta^2) = 0.02$ or 0.
- Max's payoff: $(-1) + (10.1\alpha) + (\pm\alpha^2) = 7.28$ or 8.90.

The higher value refers to the cases where Max moves to $\bot$, and the lower value refers to the cases where Max moves to $\top$ from the literal node.

The proof is in two parts. If there is a satisfying assignment, then Max can fix one and choose to go to $\top$ from all true literals and to $\bot$ from all false literals. He can then choose to go to a true literal from each clause node and to a false literal from all variable states. He can then ask Min to go left initially. As a result, all paths have payout 0 for Min, so that she does not object to going left. The payout of the maximiser is 7.28.

Vice versa, Min will only agree to go left if the payoff of all paths that are possible under a strategy selected by Max is 0 (in which case Max' α-discounted expected payout ≤ 7.28), as in all other cases her expected β-discounted payoff is > 0 for going left of < 0 when going right, so that she would favour going right over going left (resulting in an α-discounted expected payout ≤ -7.28 for Max).

To obtain this, all reachable literals from the right need to go to $\bot$, while all literals reachable from the left need to go to $\top$. This is only possible if no literal is reachable from the left *and* right. This then provides us with a satisfying assignment for our 3-SAT formula, by simply making true all literals that are reachable through the right. Note that counting does not help as the literals are always reached in exactly two steps. ◀

For inclusion in NP, we note that we can simply guess a strategy and the positional valuations for both players and check whether the valuation is consistent, meets the threshold, and ensures that Min cannot improve over following it where the policies are pure. For stochastic policies, we do not have an argument why the probabilities should be fractions with dominators in length polynomial (or: value exponential) in the size of the input. We can, however, leave them symbolic and express the constraints in the exponential theory of the reals. Combining this with the previously established hardness result, we obtain Theorem 5.

We note that this reduction works only for incentive-free equilibria. For incentive equilibria, the optimal choice for Max would be to always go to $\top$ and to pay Min an incentive of 0.02 to go left regardless, which leads to a β-discounted payoff of 0 for Min, and an α-discounted payoff of 8.88 for Max, independent of whether or not the 3-SAT formula is satisfiable.

C Complexity of Incentive Stackelberg Equilibria with Positional Strategies

This appendix refines the result of the previous appendix by showing that the problem remains hard even when incentives are allowed. In the PosStackelbergIncentive problem, the goal is to determine whether there exists a stationary incentivised profile (ι, σ, τ) such that $\mathcal{D}^{\iota,\sigma,\tau}_{\mathsf{MAX}}(s_0) \geq \lambda$. The hardness proof builds upon the earlier reduction, adapted to incorporate incentive transfers.

Proof. To prove Theorem 6 we reduce from 3-SAT via a construction that modifies the non-incentive reduction of Appendix B by allowing Player Max to offer incentives to Player Min. We construct an ASG G_ϕ from a Boolean formula ϕ such that Player Max can incentivise Player Min to take a favorable action if and only if ϕ is satisfiable.

Construction. For a CNF formula ϕ over a set $X = \{x_1, x_2, \ldots, x_n\}$ of n variables that uses the set of m clauses $C = \{C_1, C_2, \ldots, C_m\}$ and $C' = \{C'_1, C'_2, \ldots, C'_m\}$. Each clause $C_i \in C$ has only one transition to $C'_i \in C'$. Each $C'_i \in C'$ contains three literals, denoted C^1_i, C^2_i, C^3_i. We construct an ASG $G_\phi = (S, A, T, R, q_0, S_{\mathsf{MAX}}, S_{\mathsf{MIN}})$ from ϕ where:

- $S = \{q_0, q_1, \top, \bot\} \cup C \cup C' \cup \{V_x : x \in X\} \cup \{x, \neg x : x \in X\}$ is the set of states;
- $A = \{\texttt{right}, \texttt{down}\} \cup \{1, 2, 3\} \cup \{v, \neg v\} \cup \{\top, \bot\} \cup \{\star\}$;
- The transition function T is defined as:
  - $T(q_0, \texttt{right})(q_1) = 1$;
  - $T(q_1, \star, q_1) = 1$ for all $a \in A$;
  - $T(q_0, \texttt{down})(s') = \frac{1}{n+m}$ if $s' \in V_x \cup C$ and is 0 otherwise;
  - $T(s, \star)(s') = 1$ if $s \in C$ and $s' \in C'$;
  - $T(s, i)(\ell) = 1$ if $s \in C'$, $i \in \{1, 2, 3\}$ and $\ell = s^i$;
  - $T(s, v)(s') = 1$ if $s = V_x$, $s' = x$ and $T(s, \neg v)(s') = 1$ if $s = V_x$, $s' = \neg x$;
  - $T(s, \top, \top) = 1$ $T(s, \bot, \bot) = 1$ and if $s \in \{x, \neg x : x \in X\}$;
  - $T(\top, \star, \top) = 1$ and $T(\bot, \star, \bot) = 1$ for all $a \in A$;
- $S_{\mathsf{MIN}} = \{q_0\}$ and $S_{\mathsf{MAX}} = S \setminus S_{\mathsf{MIN}}$.
- $r(s, \texttt{right}) = -42$, $r(s, \top) = +3.6$, $r(s, \bot) = -1$, and $r(\top, \star) = 0$, $r(\bot, \star) = +2$. The reward for all other transitions are 0.

We let the discount factors of Max and Min are $\alpha = \frac{2}{3}$ and $\beta = \frac{1}{3}$, respectively.

Example. Consider the formula $\phi = (x \vee \neg y \vee z) \wedge (\neg x \vee \neg y \vee z) \wedge (\neg x \vee y \vee \neg z)$. A game arena G_ϕ corresponding to this formula is shown in Figure 2(b).

Proof of correctness. Again, to understand the idea behind the construction, let us first compute the values from various states for two policies of Player Min: one that chooses the `right` action in state q_0, and another that chooses `down`.

1. if Min chooses `right` action in state q_0 then all we have the following valuations:
   - MIN payoff: $0 + (-42) \cdot \frac{\beta^1}{1-\beta} = -21$.
   - MAX payoff: $0 + (-42) \cdot \frac{\alpha^1}{1-\alpha} = -84$.
2. If Min chooses `down`; Then valuations in different states are as the following: In $\top$ sink:
   - MIN value: $V^\beta_{\mathsf{MIN}}(\top) = \frac{\beta^0 \cdot 0}{1-\beta} = 0$.
   - MAX value: $V^\alpha_{\mathsf{MAX}}(\top) = \frac{\alpha^0 \cdot 0}{1-\alpha} = 0$.

   In $\bot$ sink:
   - MIN value: $V^\beta_{\mathsf{MIN}}(\bot) = \frac{\beta^0 \cdot 2}{1-\beta} = 3$.
   - MAX value: $V^\alpha_{\mathsf{MAX}}(\bot) = \frac{\alpha^0 \cdot 2}{1-\alpha} = 6$.

   In each $s \in \{x, \neg x : x \in X\}$ Max should take one of the actions from $\{\top, \bot\}$.

   If Max Takes $\top$ action:
   - MIN value: $V^\beta_{\mathsf{MIN}}(s) = r(s, \top) + \beta \cdot V^\beta_{\mathsf{MIN}}(\top) = +3.6 + \frac{1}{3} \cdot 0 = 3.6$
   - MAX value: $V^\alpha_{\mathsf{MAX}}(s) = r(s, \top) + \alpha \cdot V^\alpha_{\mathsf{MAX}}(\top) = +3.6 + \frac{2}{3} \cdot 0 = 3.6$

   If Max Takes $\bot$ action:
   - MIN value: $V^\beta_{\mathsf{MIN}}(s) = r(s, \bot) + \beta \cdot V^\beta_{\mathsf{MIN}}(\bot) = -1 + \frac{1}{3} \cdot 3 = 0$
   - MAX value: $V^\alpha_{\mathsf{MAX}}(s) = r(s, \bot) + \alpha \cdot V^\alpha_{\mathsf{MAX}}(\bot) = -1 + \frac{2}{3} \cdot 6 = 3$

Max wants Min to choose the `down` action. However, if Min does so, no resulting run can yield the same payoff she would receive by choosing `right` at q_0. If there exists a satisfying assignment, then Max can fix such an assignment and proceed as follows: for each variable, he chooses to go to $\top$ from all literals assigned true and to $\bot$ from all literals assigned false. He then routes each clause node to a true literal, and to a false literal from all variable

states. He can then incentivise Min to go `down` instead of `right`, by paying the difference between Min's payoff from going `down` compared to Min's payoff from going `right`. As a result, Min's payoff and incentive from Max to Min, will be -21 for Min, so that she does not object to going `right`. Max wants to find the best strategy in a way that keeps his payoff high, and the incentive low; The incentive that wants to pay to Min. For Max to find this strategy, he needs to keep the savings more then cost; $\frac{\mathsf{MinAdv}}{\mathsf{MaxDisAdv}} > 1$; Otherwise this is not reasonable for Max to incentivise Min. We have calculated Min and Max Payoff in the states $s \in \{x, \neg x : x \in X\}$ earlier, based on any one of the actions Max takes from $\{\top, \bot\}$.

- If Max chooses $\top$: $(V^{\beta}_{\mathsf{MIN}}(s, \top), V^{\alpha}_{\mathsf{MAX}}(s, \top)) = (3.6, 3.6)$, and
- if Max chooses $\bot$: $(V^{\beta}_{\mathsf{MIN}}(s, \bot), V^{\alpha}_{\mathsf{MAX}}(s, \bot)) = (0, 3)$

In the literals' states, $\{x, \neg x : x \in X\}$, if Max chooses to go to $\top$ sink instead of going to $\bot$ sink, then we have:

- $\mathsf{MinAdv} = V^{\beta}_{\mathsf{MIN}}(s, \top) - V^{\beta}_{\mathsf{MIN}}(s, \bot) = 3.6 - 0 = 3.6$
- $\mathsf{MaxDisAdv} = V^{\beta}_{\mathsf{MAX}}(s, \top) - V^{\beta}_{\mathsf{MAX}}(s, \bot) = 3.6 - 3 = 0.6$

As mentioned above, Max needs to have $\frac{\mathsf{MinAdv}}{\mathsf{MaxDisAdv}} > 1$. In the literals' states, $\{x, \neg x : x \in X\}$ Max has $\frac{\mathsf{MinAdv}}{\mathsf{MaxDisAdv}} = \frac{3.6}{0.6} = 6$.

Now we calculate this ratio to q_0 state. When q_0 chooses `down`, the game could stochastically go to $C = \{C_1, C_2, \ldots, C_m\}$ or $\{V_x : x \in X\}$.

- If it goes to C, the ratio would be as follow:
  - In states $\{x, \neg x : x \in X\}$, the ratio is $\frac{6}{2^0} = 6$;
  - in states $C' = \{C'_1, C'_2, \ldots, C'_m\}$, the ratio is $\frac{6}{2^1} = 3$;
  - in states $C = \{C_1, C_2, \ldots, C_m\}$, the ratio is $\frac{6}{2^2} = 1.5$; and
  - in state $\{q_0\}$, the ration is $\frac{6}{2^3} = 0.75$.
- Similarly, if it goes to V_x, the ratio would be as follow:
  - in states $\{x, \neg x : x \in X\}$, the ratio is $\frac{6}{2^0} = 6$;
  - in states $\{V_x : x \in X\}$, the ratio is $\frac{6}{2^1} = 3$; and
  - in states $\{q_0\}$, the ratio is $\frac{6}{2^2} = 1.5$.

Game starts at q_0. Considering that Max has incentivised Min to choose `down` condition to Max pays Min; Afterward two possibilities could happen:

- If game passes from C, in state $\{x, \neg x : x \in X\}$, Max should choose $\top$ over $\bot$; And Max has $\frac{\mathsf{MinAdv}}{\mathsf{MaxDisAdv}} = 0.75$. It means the ratio for choosing $\bot$ over $\top$ is 0.75; So Max's MinAdv is less then Max's $\mathsf{MaxDisAdv}$; which is not good for Max.
- If game passes from V_x, in state $\{x, \neg x : x \in X\}$, Max should choose $\bot$ over $\top$; And Max has $\frac{\mathsf{MinAdv}}{\mathsf{MaxDisAdv}} = 1.5$. It means the ratio for choosing $\bot$ over $\top$ is 1.5; So Max's MinAdv is more then Max's $\mathsf{MaxDisAdv}$; which is good for Max.

For more clarification, after Min chooses `down`, if game goes to clause nodes then all possible runs have the following valuations:

- If Max chooses $\top$ action then
  - Min's payoff: $(0) + (0\beta) + (0\beta^2) + (3.6\beta^3) + (\frac{0\beta^4}{1-\beta}) = 0.1\overline{3}$ and
  - Max's payoff: $(0) + (0\alpha) + (0\alpha^2) + (3.6\alpha^3) + (\frac{0\alpha^4}{1-\alpha}) = 1.0\overline{6}$.
- If Max chooses $\bot$ action, then
  - Min's payoff: $(0) + (0\beta) + (0\beta^2) + (-\beta^3) + (\frac{2\beta^4}{1-\beta}) = 0$ and
  - Max's payoff: $(0) + (0\alpha) + (0\alpha^2) + (-\alpha^3) + (\frac{2\alpha^4}{1-\alpha}) = 0.\overline{8}$.

  Max would have to pay a higher incentive to incentivize Min when playing $\top$, but his payoff rises by more than the difference. He would therefore play $\top$.

If the game goes to variable nodes, all possible runs have the following valuations:

- If Max chooses $\top$ action, then
  - Min's payoff: $(0) + (0\beta) + (3.6\beta^2) + (\frac{0\beta^3}{1-\beta}) = 0.4$ and
  - Max's payoff: $(0) + (0\alpha) + (3.6\alpha^2) + (\frac{0\alpha^3}{1-\alpha}) = 1.6$.
- If Max chooses $\bot$ action, then
  - Min's payoff: $(0) + (0\beta) + (-\beta^2) + (\frac{2\beta^3}{1-\beta}) = 0$ and
  - Max's payoff: $(0) + (0\alpha) + (-\alpha^2) + (\frac{2\alpha^3}{1-\alpha}) = 1.\overline{3}$.

  Max would have to pay a higher incentive to incentivize Min when playing $\top$, and his payoff rises by less than this difference. He would therefore play $\bot$.

In summary, for runs that go through clause nodes, choosing the $\top$ action is most beneficial, while for runs that go through variable nodes, choosing the $\bot$ action is most beneficial, because it decreases the payoff that needed to be paid from Max to Min as incentive: $\mathsf{benefit}_3(\ell, \bot) = (0.89 - 0) - (1.07 - 0.13) = -0.05 < 0$ and $\mathsf{benefit}_2(\ell, \bot) = (1.\overline{3} - 0) - (1.6 - 0.4) = 0.1\overline{3} > 0$ holds for every literal ℓ.

As a result, the best outcome for Max is if all reachable literals from each $V_x : x \in X$ go to $\bot$, while all literals reachable from the clauses $C = \{C_1, C_2, \ldots, C_m\}$ go to $\top$. Max will then have to pay $21 + 0.13 \cdot \frac{n}{m+n}$, so that Min still obtains her threshold outcome of -21. Max will then receive an expected payoff (before deducing the incentive he pays) of $1.\overline{3} \cdot \frac{m}{m+n} + 1.07 \cdot \frac{n}{m+n}$. After deducing the incentive he pays, this provides him with a payoff of $1.\overline{3} \cdot \frac{m}{m+n} + 0.94 \cdot \frac{n}{m+n} - 21$. This is only possible if no literal is reachable from the left *and* right. This then provides us with a satisfying assignment for our 3-SAT formula, by simply making true all literals that are reachable through the clauses. Similarly, a satisfying assignment allows us to turn to $\top$ from all literals that are true under this assignment, and to turn to a true literal from each clause node. At the same time, we can map all false literals to $\bot$ and turn to a false literal from each variable node.

The argument for inclusion in NP and ETR, respectively, is similar to the argument for inclusion in NP and ETR, respectively, for non-incentivised Stackelberg equilibria (cf. Appendix B): we can simply guess the policy explicitly for pure policies (providing inclusion in NP) and encode the constraints symbolically for stochastic policies (providing inclusion in ETR). ◀