Not Just an Empty Threat: Subgame-Perfect Equilibrium in Repeated Games Played by Computationally Bounded Players

Joseph Y. Halpern, Rafael Pass, and Lior Seeman[⋆]

Cornell University

Abstract. We study the problem of finding a subgame-perfect equilibrium in repeated games. In earlier work [Halpern, Pass and Seeman 2014], we showed how to efficiently find an (approximate) Nash equilibrium if assuming that players are computationally bounded (and making standard cryptographic hardness assumptions); in contrast, as demonstrated in the work of Borgs et al. [2010], unless we restrict to computationally bounded players, the problem is PPAD-hard. But it is well-known that for extensive-form games (such as repeated games), Nash equilibrium is a weak solution concept. In this work, we define and study an appropriate notion of a subgame-perfect equilibrium for computationally bounded players, and show how to efficiently find such an equilibrium in repeated games (again, making standard cryptographic hardness assumptions). We also show that our algorithm works not only for games with a finite number of players, but also for constant-degree graphical games.

1 Introduction

Computing a Nash equilibrium (NE) (or even an ϵ-NE) in a (one-shot) game with only two players is believed to be computationally intractable (formally, it is PPAD-Hard) [3, 4]. However, in real life, games are often played repeatedly. In infinitely repeated games, the *Folk Theorem* (see [12] for a review), which shows that the set of NE is large, gives hopes that it might be easier to find one. In two-player repeated games, Littman and Stone [11] show that this is indeed the case, and describe an efficient algorithm for finding a NE, which uses the ideas of the folk theorem. Unfortunately, Borgs et al. [2] show that if the game has three or more players, then even in the infinitely repeated version it is PPAD-Hard to find even an ϵ-NE for an inverse-polynomial ϵ.

[⋆] Joseph Halpern and Lior Seeman are supported in part by NSF grants IIS-0911036 and CCF-1214844, by AFOSR grant FA9550-08-1-0266, by ARO grant W911NF-14-1-0017, and by the Multidisciplinary University Research Initiative (MURI) program administered by the AFOSR under grant FA9550-12-1-0040. Lior Seeman is partially supported by a grant from the Simons Foundation #315783. Rafael Pass is supported in part by an Alfred P. Sloan Fellowship, a Microsoft Research Faculty Fellowship, NSF Awards CNS-1217821 and CCF-1214844, NSF CAREER Award CCF-0746990, AFOSR YIP Award FA9550-10-1-0093, and DARPA and AFRL under contract FA8750-11-2-0211. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

If we take seriously the importance of being able to find an ϵ-NE efficiently, it is partly because we have computationally bounded players in mind. But then it seems reasonable to see what happens if we assume that the players in the game are themselves computationally bounded. As we show in a recent paper [8], this makes a big difference. Specifically, we show that if we assume that players are resource bounded, which we model as probabilistic polynomial-time Turing machines (PPT TMs) with memory, and restrict the equilibrium deviations strategies to those that can be implemented by such players, then there exist an efficient algorithm for computing an ϵ-NE in an infinitely repeated game. Our equilibrium strategy uses threats and punishment much in the same way that they are used in the Folk Theorem. However, since the players are computationally bounded we can use cryptography (we assume the existence of a secure public key encryption scheme) to secretly correlate the punishing players. This allows us to overcome the difficulties raised by Borgs et al. [2].

While NE has some attractive features, it allows some unreasonable solutions. In particular, the equilibrium might be obtained by what are arguably empty threats. This actually happens in our solution (and in the basic version of the folk theorem). Specifically, players are required to punish a deviating player, even though that might hurt their payoff. Thus, if a deviation occurs, it might not be the best response of the players to follow their strategy and punish; thus, such a punishment is actually an empty threat.

To deal with this (well known) problem, a number of refinements of NE have been considered. The one typically used in dynamic games of perfect information is *subgame-perfect equilibrium*, suggested by Selten [14]. A strategy profile is a subgame-perfect equilibrium if it is a NE at every subgame of the original game. Informally, this means that at any history of the game (even those that are not on any equilibrium path), if all the players follow their strategy from that point on, then no player has an incentive to deviate. In the context of repeated games where players' moves are observed (so that it is a game of perfect information), the folk theorem continues to hold even if the solution concept used is subgame-perfect equilibrium [1, 5, 13].

In this paper, we show that we can efficiently compute a *computational subgame-perfect ϵ-equilibrium*. (The "computational" here means that we restrict deviating players to using polynomial-time strategies.) There are a number of subtleties that arise in making this precise. While we assume that all actions in the underlying repeated game are observable, we allow our TMs to also have memory, which means their action does not depend only on the public history. Like subgame-perfect equilibrium, we would like our solution concept to capture the intuition that the strategies are in equilibrium after any possible deviation. This means that in a computational subgame-perfect equilibrium, at each history for player i, player i must make a (possibly approximate) best response, no matter what his and the other players' memory states are.

Another point of view is to say that the players do not in fact have perfect information in our setting, since we allow the TMs to have memory that is not observed by the other players, and thus the game should be understood as a game of imperfect information. Subgame perfection is still defined in games of imperfect information, but in many cases does not have much bite (see [10] for a discussion on this point). For the games that we consider, subgame-perfect equilibrium typically reduces to NE. An arguably more natural generalization of subgame-perfect equilibrium in imperfect-information games would require that if an information set for player i off the equilibrium path is reached, then player i's strategy is a best response to the other players' strategies *no matter how that information set is reached*. This is quite a strong requirement. (see [12][pp. 219–221] for a discussion of this issue); such equilibria do not in general exist in games of imperfect information.[1] In our setting, a "situation" includes the players' state of memory; after a deviation, players have no idea how the state of memory of other players may have changed. Thus, the nodes in a player's information set are characterized by the possible memory states of the other players. Since in a computational subgame-perfect equilibrium, at each history for player i, player i must make a best response no matter what the memory states of the other players are, it captures the strong requirement mentioned above. Despite this, we show that in a repeated game, a computational subgame-perfect ϵ-eqilibrium exists and can be found efficiently.

To achieve this we use the same basic strategy as in [8], but, as often done to get a subgame-perfect equilibrium (for example see [5]), we limit the punishment phase length, so that the players are not incentivized not to punish deviations. However, to prove our result, we need to overcome a significant hurdle. When using cryptographic protocols, it is often the case (and, specifically is the case in the protocol used in [8]) that player i chooses a secret (e.g., a secret key for a public-key encryption scheme) as the result of some randomization, and then releases some public information which is a function of it (e.g., a public key). After that public information has been released, another party j typically has a profitable deviation by switching to the TM M that can break the protocol—for every valid public information, there always *exists* some TM M that has the secret "hardwired" into it (although there may not be an efficient way of finding M given the information). We deal with this problem by doing what is often done in practice: we do not use any key for too long, so that j cannot gain too much by knowing any one key.

A second challenge we face is that in order to prove that our proposed strategies are even an ϵ-NE, we need to show that the payoff of the *best response* to this strategy is not much greater than that of playing the strategy. However, since for any polynomial-time TM there is always a better polynomial-time TM that has just a slightly longer running time, this natural approach fails. This instead leads us to characterize a class of TMs we can analyze, and show that any other TM can be converted to a TM in this class that has at least the same payoff. While such an argument might seem simple in the traditional setting, since we only allow for polynomial time TMs, in our setting this turns out to require a surprisingly delicate construction and analysis to make sure this converted TM does indeed has the correct size and running time.

There are a few recent papers that investigate solution concepts for extensive-form games involving computationally bounded player [9, 6, 7]; some of these focus on cryptographic protocols [9, 6]. Kol and Naor [9] discuss refinements of NE in the context of cryptographic protocols, but their solution concept requires only that on each history on the equilibrium path, the strategies from that point on form a NE. Our requirement is much stronger. Gradwohl, Livne and Rosen [6] also consider this scenario and offer a solution concept different from ours; they try to define when an empty threat occurs,

[1] Indeed, that is part of why notions like *sequential equilibrium* [10] are typically considered in games of imperfect information.

and look for strategy profiles where no empty threats are made. Again, our solution concept is much stronger.

2 Preliminaries

We briefly review some relevant material regarding games and subgame-perfect equilibrium. This material largely repeats definitions from [8]. For a more detailed description and for the definition of the cryptographic primitives we use, see the full paper.

2.1 One-shot games

We define a game G to be a triple $([c], A, \vec{u})$, where $[c] = \{1, \ldots, c\}$ is the set of players, A_i is the set of possible actions for player i, $A = A_1 \times \ldots \times A_c$ is the set of action profiles, and $\vec{u} : A \rightarrow \mathbb{R}^c$ is the utility function ($\vec{u}_i(\vec{a})$ is the utility of player i). A (mixed) *strategy* σ_i for player i is a probability distribution over A_i, that is, an element of $\Delta(A_i)$ (where, as usual, we denote by $\Delta(X)$ the set of probability distributions over the set X). We use the standard notation $\vec{x}_{-i}$ to denote vector $\vec{x}$ with its ith element removed, and $(x', \vec{x}_{-i})$ to denote $\vec{x}$ with its ith element replaced by x'.

Definition 1. *(Nash Equilibrium) $\sigma = (\sigma_1, ..., \sigma_c)$ is an ϵ-NE of G if, for all players $i \in [c]$ and all actions $a'_i \in A_i$, $E_{\sigma_{-i}}[u_i(a'_i, \vec{a}_{-i})] \leq E_{\sigma}[u_i(\vec{a})] + \epsilon$.*

A *correlated strategy* of a game G is an element $\sigma \in \Delta(A)$. It is a *correlated equilibrium* if, for all players i, they have no temptation to play a different action, given that the action profile was chosen according to σ. That is, for all players i for all $a_i \in A_i$ such that $\sigma_i(a_i) > 0$, $E_{\sigma|a_i} u_i(a_i, \vec{a}_{-i}) \geq E_{\sigma|a_i} u_i(a'_i, \vec{a}_{-i})$.

Player i's minimax value in a game G is the highest payoff i can guarantee himself if the other players are trying to push his payoff as low as possible. We call the strategy i plays in this case a minimax strategy for i; the strategy that the other players use is i's (correlated) punishment strategy. (Of course, there could be more than one minimax strategy or punishment strategy for player i.) Note that a correlated punishment strategy can be computed using linear programming.

Definition 2. *Given a game $G = ([c], A, \vec{u})$, the strategies $\vec{\sigma}_{-i} \in \Delta(A_{-i})$ that minimize* $\max_{\sigma' \in \Delta(A_i)} E_{(\sigma', \vec{\sigma}_{-i})}[u_i(\vec{a})]$ *are the* punishment strategies *against player i in G. If $\vec{\sigma}_{-i}$ is a punishment strategy against player i, then $mm_i(G) =$* $\max_{a \in A_i} E_{\vec{\sigma}_{-i}}[u_i(a, a_{-i})]$ *is player i's* minimax value *in G*

To simplify the presentation, we assume all payoffs are normalized so that each player's minimax value is 0. Since, in an equilibrium, all players get at least their minimax value, this guarantees that all players get at least 0 in a NE.

2.2 Infinitely repeated games

Given a normal-form game G, we define the repeated game $G^t(\delta)$ as the game in which G is played repeatedly t times (in this context, G is called the *stage game*) and $1 - \delta$

is the discount factor (see below). Let $G^{\infty}(\delta)$ be the game where G is played infinitely many times. An infinite history h in this game is an infinite sequence $\langle \vec{a}^0, \vec{a}^1, \ldots \rangle$ of action profiles. Intuitively, we can think of $\vec{a}^t$ as the action profile played in the t^{th} stage game. We often omit the δ in $G^{\infty}(\delta)$ if it is not relevant to the discussion. Let $H_{G^{\infty}}$ be the set of all possible histories of G^{∞}. For a history $h \in H_{G^{\infty}}$ let $G^{\infty}(h)$ the subgame that starts at history h (after $|h|$ one-shot games have been played where all players played according to h). We assume that G^{∞} is *fully observable*, in the sense that, after each stage game, the players observe exactly what actions the other players played.

A (behavioral) strategy for player i in a repeated game is a function σ from histories of the games to $\Delta(A_i)$. Note that a profile $\vec{\sigma}$ induces a distribution $\rho_{\vec{\sigma}}$ on infinite histories of play. Let $\rho^t_{\vec{\sigma}}$ denote the induced distribution on H^t, the set of histories of length t. (If $t = 0$, we take H^0 to consist of the unique history of length 0, namely $\langle\,\rangle$.) Player i's utility if $\vec{\sigma}$ is played, denoted $p_i(\vec{\sigma})$, is defined as follows:

$$p_i(\vec{\sigma}) = \delta \sum_{t=0}^{\infty} (1-\delta)^t \sum_{h \in H^t, \vec{a} \in A} \rho^{t+1}_{\vec{\sigma}}(h \cdot \vec{a})[u_i(\vec{a})].$$

Thus, the discount factor is $1 - \delta$. Note that the initial δ is a normalization factor. It guarantees that if $u_i(\vec{a}) \in [b_1, b_2]$ for all joint actions $\vec{a}$ in G, then i's utility is in $[b_1, b_2]$, no matter which strategy profile $\vec{\sigma}$ is played.

In these game, a more robust solution concept is subgame-perfect equilibrium [14], which requires that the strategies form an ϵ-NE at every history of the game.

Definition 3. *A strategy profile $\vec{\sigma} = (\sigma_1, ..., \sigma_c)$, is a subgame-perfect ϵ-equilibrium of a repeated game G^{∞}, if, for all players $i \in [c]$, all histories $h \in H_{G^{\infty}}$ where player i moves, and all strategies σ' for player i, $p^h_i((\sigma')^h, \vec{\sigma}^h_{-i}) \leq p^h_i(\vec{\sigma}^h) + \epsilon$, where p^h_i is the utility function for player i in game $G^{\infty}(h)$, and σ^h is the restriction of σ to $G^{\infty}(h)$.*

3 Computational subgame-perfect equilibrium

In this section we define our solution concept, and show that it can be computed efficiently in a repeated game. We capture computational boundedness by considering only (possibly probabilistic) polynomial time TMs, which at round t use only polynomial in nt many steps to compute the next action, where n is the size of G (the max of the number of actions and the number of players in G). The TM gets as input the history of play so far and can also use internal memory that persists from round to round. A strategy for player i is then a TM M_i. Given a strategy profile $\vec{M}$, as above, we can define the induced distribution $\rho_{\vec{M}}$ and player i's payoff $p_i(\vec{M})$.

We would like to define a notion similar to subgame-perfect equilibrium, where for all histories h in the game tree (even ones not on the equilibrium path), playing $\vec{\sigma}$ restricted to the subtree starting at h forms a NE. This means that a player does not have any incentive to deviate, no matter where he finds himself in the game tree.

As we suggested in the introduction, there are a number of issues that need to be addressed in formalizing this intuition in our computational setting. First, since we consider stateful TMs, there is more to a description of a situation than just the history;

we need to know the memory state of the TM. That is, if we take a history to be just a sequence of actions, then the analogue of history for us is really a pair $(h, \vec{m})$ consisting of a sequence h of actions, and a profile of memory states, one for each player. Thus, to be a computational subgame-perfect equilibrium the strategies should be a NE at every history and no matter what the memory states are.

Since a player's TM cannot observe the memory state of the other players' TMs, the computational game is best thought of as a game of imperfect information, where, in a given history h where i moves, i's information set consists of all situations where the history is h and the states of memory of the other players are arbitrary. While subgame-perfect equilibrium extends to imperfect information games it usually doesn't have much bite. In our setting it reduces to just a NE.

Instead, in games of imperfect information, the solution concept most commonly used is *sequential equilibrium* [10]. A sequential equilibrium is a pair $(\vec{\sigma}, \mu)$ consisting of a strategy profile $\vec{\sigma}$ and a *belief system* μ, where μ associates with each information set I a probability $\mu(I)$ on the nodes in I. Intuitively, if I is an information set for player i, $\mu(I)$ describes i's beliefs about the likelihood of being in each of the nodes in I. Then $(\vec{\sigma}, \mu)$ is a sequential equilibrium if, for each player i and each information set I for player i, σ_i is a best response to $\vec{\sigma}_{-i}$ given i's beliefs $\mu(I)$. However, a common criticism of this solution concept is that it is unclear what these beliefs should be and how players create these beliefs. Instead, our notion of computational subgame-perfection can be viewed as a strong version of a sequential equilibrium, where, for each player i and each information set I for i, σ_i is a best response to $\vec{\sigma}_{-i}$ conditional on reaching I (up to ϵ) no matter what i's beliefs are at I.

As a deviating TM can change its memory state in arbitrary ways, when we argue that a strategy profile is an ϵ-NE at a history, we must also consider all possible states that the TM might start with at that history. Since there exists a deviation that just rewrites the memory in the round just before the history we are considering, any memory state (of polynomial length) is possible. Thus, in the computational setting, we require that the TM's strategies are an ϵ-NE at every history, no matter what the states of the TMs are at that history. This solution concept is in the spirit of subgame-perfect equilibrium, as we require that the strategies are a NE after every possible deviation, although the player might not have complete information as to what the deviation is.

Intuitively, a profile $\vec{M}$ of TMs is a computational subgame-perfect equilibrium if for all players i, all histories h where i moves, and all memory profiles $\vec{m}$ of the players, there is no polynomial-time TM $\bar{M}$ such that player i can gain more than ϵ by switching from M_i to $\bar{M}$. Unfortunately, what we have just said is meaningless if we consider only a single game. The notion of "polynomial-time TM" does not make sense for a single game. To make it precise, we must consider an infinite sequence of games of increasing size (just as was done in [8], although our current definition is more complicated since we must consider memory states).

For a memory state m and a TM M let $M(m)$, stand for running M with initial memory state m. We use $\vec{M}(\vec{m})$ to denote $(M_1(m_1), \dots, M_c(m_c))$. Let $p_i^{G,\delta}(\vec{M})$ denote player i's payoff in $G^{\infty}(\delta)$ when $\vec{M}$ is played.

Definition 4. *An infinite sequence of strategy profiles $\vec{M}^1, \vec{M}^2, \ldots$, where $\vec{M}^k = (M_1^k, ..., M_c^k)$, is a* computational subgame-perfect ϵ-equilibrium *of an* *infinite sequence $G_1^\infty, G_2^\infty, \ldots$ of repeated games where the size of G_k is k, if, for all players $i \in [c]$, all sequences $h_1 \in H_{G_1^\infty}, h_2 \in H_{G_2^\infty}, \ldots$ of histories, all sequences $\vec{m}^1, \vec{m}^2, \ldots$ of polynomial-length memory-state profiles, where $\vec{m}^k = (m_1^k, \ldots, m_c^k)$, and all non-uniform PPT adversaries $\bar{M}$ (polynomial in k and t, as discussed above), there exists k_0 such that, for all $k \geq k_0$,*

$$p_i^{G_k^\infty(h_k),\delta}(\bar{M}(m_i^k), \vec{M}_{-i}^k(\vec{m}_{-i}^k)) \leq p_i^{G_k^\infty(h_k),\delta}(\vec{M}^k(\vec{m}^k)) + \epsilon(k).$$

We stress that our equilibrium notion considers only deviations that can be implemented by polynomial-time TMs. This differs from the standard definition of NE (and from the definition considered in [2]). But this difference is exactly what allows us to use cryptographic techniques. It is also the reason that we need to consider a sequence of games of growing sizes instead of a single game. We allow the deviation to be a *non-uniform PPT*, which can be viewed as a sequence of TMs whose running time is bounded by some common polynomial (see the full paper for a formal definition).

3.1 Computing a subgame-perfect ϵ-NE

Let $A_i^0 \subset A_i$ be a non-empty set and let $A_i^1 = A_i \setminus A_i^0$.[2] A player can broadcast an m-bit string by using his actions for m rounds, by treating actions from A_i^0 as 0 and actions from A_i^1 as 1. Given a polynomial ϕ (with natural coefficients), let $(\mathrm{Gen}, \mathrm{Enc}, \mathrm{Dec})$ be a multi-message multi-key secure ϕ-bit (see the full paper for the definition), if the security parameter is k, the length of an encrypted message is $z(k)$ for some polynomial z. Let $sq = (s_1, s_2 \ldots, s_m)$ be a fixed sequence of action profiles. Fix a polynomial-time pseudorandom function ensemble $\{PS_s : s \in \{0,1\}^*\}$ (again, see the full paper for the definition).

For a game G such that $|G| = n$, and a polynomial ℓ, consider the following strategy $\sigma^{NE,\ell}$, and let $\vec{M}^{\sigma^{NE,\ell}}$ be the TM that implements this strategy. This strategy is similar in spirit to that proposed in [8]; indeed, the first two phases are identical. Phase 1 explains what to do if no deviation occurs: play sq. Phase 2 gives the preliminaries of what to do if a deviation does occur: roughly, compute a random seed that is shared with all the non-deviating players. Phase 3 explains how to use the random seed to produce a correlated punishment strategy that punishes the deviating player. The key difference between the strategy here and that in [8] is that this punishment phase is played for only $\ell(n)$ rounds. After that, players return to phase 1. As we show, this limited punishment is effective since it is not played long enough to make it an empty threat (if ℓ is chosen appropriately). Phase 4 takes care of one minor issue: The fact that we can start in any memory state means that a player might be called on to do something that, in fact, he cannot do (because he doesn't have the information required to do it). For example, he might be called upon to play the correlated punishment strategy in a state where he has forgotten the random seed, so he cannot play it. In this case, a default action is played.

1. Play according to sq (with wraparound) as long as all players played according to sq in the previous round.

[2] We assume that each player has at least two actions in G. This assumption is without loss of generality—we can essentially ignore players for whom it does not hold.

2. After detecting a deviation by player $j \neq i$ in round t_0:[3]
   (a) Generate a pair (pk_i, sk_i) using Gen(1^n). Store sk_i in memory and use the next $l(n)$ rounds to broadcast pk_i, as discussed above.
   (b) If $i = j + 1$ (with wraparound), player i does the following:
      - i records $pk_{j'}$ for all players $j' \notin \{i, j\}$;
      - i generates a random n-bit seed $seed$;
      - for each player $j' \notin \{i, j\}$, i computes $m = Enc_{pk_{j'}}(seed)$, and uses the next $(c - 2)z(n)$ rounds to communicate these strings to the players other than i and j (in some predefined order).
   (c) If $i \neq j + 1$, player i does the following:
      - i records the actions played by $j+1$ at time slots designated for i to retrieve $Enc_{Pk_i}(seed)$;
      - i decrypts to obtain $seed$, using Dec and sk_i.
3. Phase 2 ends after $\phi(n)+(c-2)z(n)$ rounds. The players other than j then compute $PS_{seed}(t)$ and use it to determine which action profile to play according to the distribution defined by a fixed (correlated) punishment strategy against j. Player j plays his best response to the correlated punishment strategy throughout this phase. After $\ell(n)$ rounds, they return to phase 1, playing the sequence sq from the point at which the deviation occurred (which can easily be inferred from the history).
4. If at any point less than or equal to $\phi(n) + (c - 2)z(n)$ time steps from the last deviation from phase 1 the situation is incompatible with phase 2 as described above (perhaps because further deviations have occurred), or at any point between $\phi(n) + (c - 2)z(n)$ and $\phi(n) + (c - 2)z(n) + \ell(n)$ steps since the last deviation from phase 1 the situation is incompatible with phase 3 as described above, play a fixed action for the number of rounds left to complete phases 2 and 3 (i.e., up to $\phi(n) + (c-2)z(n) + \ell(n)$ steps from the last deviation from phase 1). Then return to phase 1.

Note that with this strategy a deviation made during the punishment phase is not punished. Phase 2 and 3 are always played to their full length (which is fixed and predefined by ℓ and z). We say that a history h is a phase 1 history if it is a history where an honest player should play according to sq. History h is a phase 2 history if it is a history where at most $\phi(n)+(c-2)z(n)$ rounds have passed since the last deviation from phase 1; h is a phase 3 history if more than $\phi(n)+(c-2)z(n)$ but at most $\phi(n)+(c-2)z(n)+\ell(n)$ rounds have passed since the last deviation from phase 1. No matter what happens in phase 2 and 3, a history in which exactly $\phi(n) + (c-2)z(n) + \ell(n)$ round have passed since the last deviation from phase 1 is also a phase 1 history (even if the players deviate from phase 2 and 3 in arbitrary ways). Thus, no matter how many deviations occur, we can uniquely identify the phase of each round.

We next show that by selecting the right parameters, these strategies are easy to compute and are a subgame-perfect ϵ-equilibrium for all inverse polynomials ϵ.

[3] If more than one player deviates while playing sq, the players punish the one with the smaller index. The punished player plays his best response to what the other players are doing in this phase.

Definition 5. *Let $\mathcal{G}_{a,b,c,n}$ be the set of all games with c players, at most n actions per player, integral payoffs,*[4] *maximum payoff a, and minimum payoff b.*[5]

Our proof uses the following three lemmas proved in [8]. The first lemma shows that, given a correlated strategy σ in a game G, players can get an average payoff that is arbitrarily close to their payoff in σ by playing a fixed sequence of action profiles repeatedly.

Lemma 1. *For all a, b, c, all polynomials q, all n, all games $G \in \mathcal{G}_{a,b,c,n}$, and all correlated strategies σ in G, if the expected payoff vector of playing σ is ξ, then there exists a sequence sq of action profiles of length $w(n) = 2((a-b)q(n)+1)n^c$, such that for all $\delta \leq 1/f(n)$, where $f(n) = 2(a-b)w(n)q(n)$, if sq is played infinitely often, then player i's payoff in $G^\infty(\delta)$ is at least $\xi_i - 1/q(n)$, no matter at which point of the sequence play is started.*

To explain what we mean by the phrase "no matter at which point of the sequence play is started", suppose that the sequence sq has the form $(\vec{a}_1, \ldots, \vec{a}_5)$. Then the result holds if we start by playing $\vec{a}_1$ or if we start by playing $\vec{a}_4$ (and then continue with $\vec{a}_5$, $\vec{a}_1$, $\vec{a}_2$, and so on).

The next lemma shows that, for every inverse polynomial, if we "cut off" the game after some appropriately large polynomial p number of rounds (and compute the discounted utility for the finitely repeated game considering only $p(n)$ repetitions), the difference between a player's utility in the infinitely repeated and the finitely repeated game is negligible; that is, the finitely repeated game is a "good" approximation of the infinitely repeated game.

Lemma 2. *For all a, b, c, all polynomials q, all n, all games $G \in \mathcal{G}_{a,b,c,n}$, all $0 < \delta < 1$, all strategy profiles $\vec{M}$, and all players i, the difference between i's expected utility $p_i[\vec{M}]$ in game $G^{\lceil n/\delta \rceil}(\delta)$ and $p_i[\vec{M}]$ in game $G^\infty(\delta)$ is at most $a/2^n$.*

The last lemma shows that the punished player's expected payoff is negligible.

Lemma 3. *For all a, b, c, all polynomials t and f, all sequences of games $G_1, G_2, \ldots$ such that $G_n \in \mathcal{G}_{a,b,c,n}$, and all players i, if the players other than i play $\vec{M}_{-i}^{\sigma^{NE,\ell}}$, then for all non-uniform polynomial time TMs M, there exists a negligible function ϵ such that if i uses M and M deviates from the phase 1 sequence before round $t(n)$, then i's expected payoff during phase 3 is less then $\epsilon(n)$.*

We first show that for any strategy that deviates while phase 1 is played, there is a strategy whose payoff is at least as good and either does not deviate in the first polynomially many rounds, or after its first deviation, deviates every time phase 1 is played. (Recall that after every deviation in phase 1, the other players play the punishment phase for $\ell(n)$ rounds and then play phase 1 again.)

[4] Our result also hold for rational payoffs, except then the size of the game needs to take into account the bits needed to represent the payoffs.

[5] By our assumption that the minimax payoff is 0 for all players, we can assume $a \geq 0$, $b \leq 0$, and $a - b > 0$ (otherwise $a = b = 0$, which makes the game uninteresting).

We do this by showing that if player i has a profitable deviation at some round t of phase 1, then it must be the case that every time this round of phase 1 is played, i has a profitable deviation there. (That is, the strategy of deviating every time this round of phase 1 is played is at least as good as a strategy where player i correlates his plays in different instantiations of phase 1.) While this is trivial in traditional game-theoretic analyses, naively applying it in the computational setting does not necessarily work. It requires us to formally show how we reduce a polynomial time TM M to a different TM M' of the desired form without blowing up the running time and size of the TM.

For a game G, let $H_{G^\infty}^{1,n,f}$ be the set of histories h of G^∞ of length at most $nf(n)$ such that at (the last node of) h, $\sigma^{NE,\ell}$ is in phase 1. Let $R(M)$ be the polynomial that bounds the running time of TM M.

Definition 6. *Given a game G, a deterministic TM M is said to be (G,f,n)-*well-behaved *if, when $(M, \sigma_{-i}^{NE,\ell})$ is played, then either M does not deviate for the first $nf(n)$ rounds or, after M first deviates, M continues to deviate from sq every time phase 1 is played in the next $nf(n)$ rounds.*

Lemma 4. *For all a, b, c, and all polynomials f, there exists a polynomial g such that for all n, all games $G \in \mathcal{G}_{a,b,c,n}$, all $h \in H_{G^\infty}^{1,n,f}$, all players i, and all TMs M, there exists a $(G(h),f,n)$-well-behaved TM M' such that $p_i^{G^h,1/f(n)}(M', \vec{M}^{\sigma_{-i}^{NE,\ell}}) \geq p_i^{G^h,1/f(n)}(M, \vec{M}^{\sigma_{-i}^{NE,\ell}})$, and $R(M'), |M'| \leq g(R(M))$.*

Proof. Suppose that we are given $G \in \mathcal{G}_{a,b,c,n}$, $h \in H_{G^\infty}^{1,n,f}$, and a TM M. We can assume without loss of generality that M is deterministic (we can always just use the best random tape). If M does not deviate in the first $nf(n)$ rounds of $G(h)^\infty$ then M' is just M, and we are done. Otherwise, we construct a sequence of TMs starting with M that are, in a precise sense, more and more well behaved, until eventually we get the desired TM M'.

For $t_1 < t_2$, say that M is (t_1,t_2)-(G,f,n)-*well-behaved* if M does not deviate from sq until round t_1, and then deviates from sq every time phase 1 is played up to (but not including) round t_2 (by which we mean there exists some history in which M does not deviate at round t_2 and this is the shortest such history over all possible random tapes of $\vec{M}^{\sigma_{-i}^{NE,\ell}}$). We construct a sequence $M_1, M_2, \ldots$ of TMs such that (a) $M_1 = M$, (b) M_i is (t_1^i, t_2^i)-(G,f,n)-well-behaved, (c) either $t_1^{i+1} > t_i$ or $t_1^{i+1} = t_1^i$ and $t_2^{i+1} > t_2^i$, and (d) $p_i^{G^h,1/f(n)}(M_{i+1}, \vec{M}^{\sigma_{-i}^{NE,\ell}}) \geq p_i^{G^h,1/f(n)}(M_i, \vec{M}^{\sigma_{-i}^{NE,\ell}})$. Note that if $t_1 \geq nf(n)$ or $t_2 \geq t_1 + nf(n)$, then a (t_1,t_2)-(G,f,n)-well-behaved TM is (G,f,n)-well-behaved.

Let $t < nf(n)$ be the first round at which M deviates. (This is well defined since the play up to t is deterministic.) Let the history up to time t be h^t. If M deviates every time that phase 1 is played for the $nf(n)$ rounds after round t, then again we can take $M' = M$, and we are done. If not, let t' be the first round after t at which phase 1 is played and there exists some history of length t' at which M does not deviate. By definition, M is (t,t')-(G,f,n)-well behaved. We take $M_1 = M$ and $(t_1^1, t_2^1) = (t,t')$. (Note that since $\vec{M}^{\sigma_{-i}^{NE,\ell}}$ are randomized during phase 2, the first time after t at which M returns to playing phase 1 and does not deviate may depend on the results of their coin tosses. We take t' to be the first time this happens with positive probability.)

Let s^{h^*} be M's memory state at a history h^*. We assume for ease of exposition that M encodes the history in its memory state. (This can be done, since the memory state at time t is of size polynomial in t.) Consider the TM M'' that acts like M up to round t, and copies M's memory state at that round (i.e., s^{h^t}). M'' continues to plays like M up to the first round t' with $t < t' < t + nf(n)$ at which $\sigma^{NE,\ell}$ would be about to return to phase 1 and M does not deviate (which means that M plays an action in the sequence sq at round t'). At round t', M'' sets its state to s^{h^t} and simulates M from history h^t with states $s^{h(t)}$; so, in particular, M'' does deviate at time t'. (Again, the time t' may depend on random choices made by $\vec{M}^{\sigma^{NE,\ell}}_{-i}$. We assume that M'' deviates the first time M is about to play phase 1 after round t and does not deviate, no matter what the outcome of the coin tosses.) This means, in particular, that M'' deviates at any such t'. We call M'' a *type 1 deviation from M*.

If $p_i^{G^{h^t},1/f(n)}(M'', \vec{M}^{\sigma^{NE,\ell}}_{-i}) > p_i^{G^{h^t},1/f(n)}(M, \vec{M}^{\sigma^{NE,\ell}}_{-i})$, then we take $M_2 = M''$. Note that $t_1^2 = t_1^1 = t$, while $t_2^2 > t_2^1 = t'$, since M'' deviates at t'. If $p_i^{G^{h^t},1/f(n)}(M'', \vec{M}^{\sigma^{NE,\ell}}_{-i}) < p_i^{G^{h^t},1/f(n)}(M, \vec{M}^{\sigma^{NE,\ell}}_{-i})$, then there exists some history h^* of both M and M'' such that $t < |h^*| < t + nf(n)$, M'' deviates at h^*, M does not, and M has a better expected payoff than M'' at h^*. (This is a history where the type 1 deviation failed to improve the payoff.) Take M_2 to be the TM that plays like $\vec{M}^{\sigma^{NE,\ell}}_{i}$ up to time t, then sets its state to s^{h^*}, and then plays like M with state s^{h^*} in history h^*. We call M_2 a *type 2 deviation from M*. Note that M_2 does not deviate at h^t (since M did not deviate at history h^*). Let $\delta' = (1-\delta)^{|h^*|-|h^t|}$. Clearly $\delta' p_i^{G^{h^t},1/f(n)}(M_2, \vec{M}^{\sigma^{NE,\ell}}_{-i}) = p_i^{G^{h^*},1/f(n)}(M, \vec{M}^{\sigma^{NE,\ell}}_{-i})$, since $\vec{M}^{\sigma^{NE,\ell}}_{-i}$ acts the same in G^{h^t} and G^{h^*}. Since M'' plays like $M(s^{h_t})$ at h^*, $p_i^{G^{h^*},1/f(n)}(M'', \vec{M}^{\sigma^{NE,\ell}}_{-i}) = \delta' p_i^{G^{h^t},1/f(n)}(M, \vec{M}^{\sigma^{NE,\ell}}_{-i})$. Combining this with the previous observations, we get that $p_i^{G^{h^t},1/f(n)}(M_2, \vec{M}^{\sigma^{NE,\ell}}_{-i}) \geq p_i^{G^{h^t},1/f(n)}(M, \vec{M}^{\sigma^{NE,\ell}}_{-i})$. Also note that $t_1^2 > t_1^1$. This completes the construction of M_2. We inductively construct M_{i+1}, $i = 2, 3, \ldots$, just as we did M_2, letting M_i play the role of M.

Next observe that, without loss of generality, we can assume that this sequence arises from a sequnce of type 2 deviations, followed by a sequence of type 1 deviations: For let j_1 be the first point in the sequence at which a type 1 deviation is made. We claim that we can assume without loss of generality that all further deviations are type 1 deviations. By assumption, since M_{j_1} gives i higher utility than M_{j_1-1}, it is better to deviate the first time M_{j_1-1} wants to play phase 1 again after an initial deviation. This means that when M_{j_1} wants to play phase 1 again after an initial deviation it must be better to deviate again, since the future play of the $\vec{M}^{\sigma^{NE,\ell}}_{-i}$ is the same in both of these situations. This means that once a type 1 deviation occurs, we can assume that all further deviations are type 1 deviations.

Let M_j be the first TM in the sequence that is well behaved. (As we observed earlier, there must be such a TM.) Using the fact that the sequence consists of a sequence of type 2 deviations followed by a sequence of type 1 deviations, it is not hard to show that M_j can be implemented efficiently. First notice that M_{j_1} is a TM that plays like $\vec{M}^{\sigma^{NE,\ell}}_{i}$ until some round, and then plays M starting with its state at a history which is at most $(nf(n))^2$ longer than the real history at this point. This is because its initial history becomes longer by at most $nf(n)$ at each round and we iterate this construction at most $nf(n)$ times. This means that its running time is obviously polynomially related to the running time of the original M. The same is true of the size of M_{j_1}, since we need to encode only the state at this initial history and the history at which we switch, which is polynomially related to $R(M)(n)$.

To construct M_j, we need to modify M_{j_1} only slightly, since only type 1 deviations occur. Specifically, we need to know only $t^1_{j_1}$ and to encode its state at this round. At every history after that, we run M_{J_1} (which is essentially running M on a longer history) on a fixed history, with a potential additional step of copying the state. It is easy to see that the resulting TM has running time and size at most $O(R(M))$. □

We now state and prove our theorem, which shows that there exists a polynomial-time algorithm for computing a subgame-perfect ϵ-equilibrium by showing that, for all inverse polynomials ϵ, there exists a polynomial function ℓ of ϵ such that $\sigma^{NE^*,\ell}$ is a subgame-perfect ϵ-equilibrium of the game. The main idea of the proof is to show that the players can't gain much from deviating while the sequence is being played, and also that, since the punishment is relatively short, deviating while a player is being punished is also not very profitable.

Theorem 1. *For all a, b, c, and all polynomials q, there is a polynomial f and a polynomial-time algorithm F such that, for all sequences $G_1, G_2, \ldots$ of games with $G^j \in G_{a,b,c,j}$ and for all inverse polynomials $\delta \leq 1/f$, the sequence of outputs of F given the sequence $G_1, G_2, \ldots$ of inputs is a subgame-perfect $\frac{1}{q}$-equilibrium for $G_1^\infty(\delta_1), G_2^\infty(\delta_2), \ldots$.*

Proof. Given a game $G^n \in \mathcal{G}(a, b, c, n)$, the algorithm finds a correlated equilibrium σ of G^n, which can be done in polynomial time using linear programming. Each player's expected payoff is at least 0 when playing σ, since we assumed that the minimax value of the game is 0. Let $r = a - b$. By Lemma 1, we can construct a sequence sq of length $w(n) = 4(rnq(n) + 1)n^c$ and set $f'(n) = 4rw(n)q(n)$, so that if the players play sq infinitely often and $\delta < 1/f'(n)$, then all the players get at least $-1/2q(n)$. The correlated punishment strategy against each player can also be found in polynomial time using linear programming.

Let $m(n)$ be the length of phase 2, including the round where the deviation occurred. (Note that $m(n)$ is a polynomial that depends only on the choice of encryption scheme—that is, it depends on ϕ, where a ϕ-bit public-key encryption scheme is used, and on z, where $z(k)$ is the length of encrypted messages.) Let $\ell(n) = nq(n)(m(n)a + 1)$, let σ_n^* be the strategy $\vec{M}^{\sigma^{NE,\ell}}$ described above, and let $f(n) = \max(3rq(n)(\ell(n) + m(n)), f'(n))$.

We now show that $\sigma_1^*, \sigma_2^*, \ldots$ is a subgame-perfect $(1/q)$-equilibrium for every inverse polynomial discount factor $\delta \leq 1/f$. We focus on deviations at histories of length $< \frac{n}{\delta(n)}$, since, by Lemma 2, the sum of payoffs received after that is negligible. Thus, there exists some n_0 such that, for all $n > n_0$, the payoff achieved after that history is less than $1/q(n)$, which does not justify deviating.

We first show that no player has an incentive to deviate in subgames starting from phase 1 histories. By Lemma 4, it suffices to consider only a deviating strategy that after its first deviation deviates every time phase 1 is played; for every deviating strategy, either not deviating does at least as well or there is a deviating strategy of this form that does at least as well. Let h_1 be the history in which the deviation occurs and let M be the deviating strategy. Notice that $\vec{M}^{\sigma^{NE,\ell}}$ can always act as intended at such histories; it can detect it is in such a history and can use the history to compute the next move (i.e., it does not need to maintain memory to figure out what to do next).

The player's payoff from $(M, \vec{M}^{\sigma^{NE,\ell}}_{-i})$ during one cycle of deviation and punishment can be at most a at each round of phase 2 and, by Lemma 3, is negligible throughout phase 3. (We use ϵ_{neg} to denote the negligible payoff to a deviator in phase 3.) Thus, the payoff of the deviating player from $(M, \vec{M}^{\sigma^{NE,\ell}}_{-i})$ from the point of deviation onwards is at most

$$((1-\delta(n)^{|h_1|})\big(\delta(n)(m(n)a+\epsilon_{neg})\sum_{t=0}^{\lceil \frac{nf(n)-|h_1|}{m(n)+\ell(n)} \rceil}(1-\delta(n))^{(m(n)+\ell(n))t}+\epsilon'_{neg}\big)$$
$$\leq ((1-\delta(n)^{|h_1|})\big(\delta(n)(m(n)a+\epsilon_{neg})\sum_{t=0}^{\infty}(1-\delta(n))^{(m(n)+\ell(n))t}+\epsilon'_{neg}\big),$$

where ϵ'_{neg} is the expected payoff after round $nf(n)$. By Lemma 1, no matter where in the sequence the players are, the average discounted payoff at that point from playing honestly is at least $-1/2q(n)$. Thus, the payoff from playing $(\vec{M}^{\sigma^{NE,\ell}})$ from this point onwards is at least $-(1-\delta(n))^{|h_1|})1/2q(n)$. We can ignore any payoff before the deviation since it is the same whether or not the player deviates. and also divide both sides by $(1-\delta(n))^{|h_1|})$; thus, it suffices to prove that

$$\delta(n)(m(n)a+\epsilon_{neg})\sum_{t=0}^{\infty}(1-\delta(n))^{(m(n)+\ell(n))t}+\epsilon'_{neg}\leq\frac{1}{2q(n)}.$$

The term on the left side is bounded by $O\big(\frac{m(n)a+\epsilon_{neg}}{nq(n)(m(n)a+1)}\big)$, and thus there exists n_1 such that, for all $n > n_1$, the term on the left side is smaller than $\frac{1}{2q(n)}$ (In fact, for all constants c, there exists n_c such that the left-hand side is at most $\frac{1}{cq(n)}$ for any $n > n_c$.)

We next show that no player wants to deviate in phase 2 or 3 histories. Notice that since these phases are carried out to completion even if the players deviate while in these phases (we do not punish them for that), and the honest strategy can easily detect whether it is in such a phase by looking at when the last deviation from phase 1 occurred. First consider a punishing player. By not following the strategy, he can gain at most r for at most $\ell(n)+m(n)$ rounds over the payoff he gets with the original strategy (this is true even if his memory state is such that he just plays a fixed action, or even if another player deviates while the phase is played). Once the players start playing phase 1 again, our previous claim shows that no matter what the actual history is at that point, a strategy that does not follow the sequence does not gain much. It is easy to verify that, given the discount factor, a deviation can increase his discounted payoff by at most $\frac{1}{q(n)}$ in this case. (Notice that the previous claim works for any constant fraction of $1/q(n)$, which is what we are using here since the deviation in the punishment phase gains $1/cq(n)$ for some c.)

The punished player can deviate to a TM that correctly guessed the keys chosen (or the current TM's memory state might contain the actual keys and he defects to a TM that uses these keys) , in which case he would know exactly what the players are going to do while they are punishing him. Such a deviation exists once the keys have been played and are part of the history. Another deviation might be a result of the other TMs being in an inconsistent memory state, so that they play a fixed action, one which the deviating player might be able to take advantage of. However, these deviations work (or any other possible deviation) only for the current punishment phase. Once the players go back to playing phase 1, this player can not gain much by deviating from the sequence again. For if he deviates again, the other players will choose new random keys and a new random seed (and will have a consistent memory state); from our previous claims, this means that no strategy can gain much over a strategy that follows the sequence. Moreover, he can also gain at most r for at most $\ell(n)+m(n)$ rounds which, as claimed before, means that his discounted payoff difference is less than $\frac{1}{q(n)}$ in this case.

This shows that, for n sufficiently large, no player can gain more than $1/q(n)$ from deviating at any history. Thus, this strategy is a subgame-perfect $1/q$-equilibrium. □

References

[1] R. J. Aumann and L. S. Shapley. Long-term competition: a game-theoretic analysis. In N. Megiddo, editor, *Essays in Game Theory*, pages 1–15. Springer, 1994.

[2] C. Borgs, J. T. Chayes, N. Immorlica, A. T. Kalai, V. S. Mirrokni, and C. H. Papadimitriou. The myth of the folk theorem. *Games and Economic Behavior*, 70(1):34–43, 2010.

[3] X. Chen, X. Deng, and S.-H. Teng. Settling the complexity of two-player Nash equilibrium. *Journal of the ACM*, 53(3), 2009.

[4] C. Daskalis, P. W. Goldberg, and C. H. Papadimitriou. The complexity of computing a Nash equilibrium. In *Proc. 38th ACM Symp. Theory of Computing*, pages 71–78, 2006.

[5] D. Fudenberg and E. Maskin. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554, 1986.

[6] R. Gradwohl, N. Livne, and A. Rosen. Sequential rationality in cryptographic protocols. *ACM Trans. Econ. Comput.*, 1(1):2:1–2:38, January 2013.

[7] J. Y. Halpern and R. Pass. Sequential equilibrium in computational games. In *Proc. 23rd International Joint Conf. Artificial Intelligence (IJCAI '13)*, pages 171–176, 2013.

[8] J.Y. Halpern, R. Pass, and L. Seeman. The truth behind the myth of the folk theorem. In *Proc. 5th Conf. Innovations in Theoretical Computer Science (ITCS '14)*, pages 543–554, 2014.

[9] G. Kol and M. Naor. Games for exchanging information. In *Proc. 40th Annual ACM Symp. on Theory of Computing (STOC '08)*, pages 423–432, 2008.

[10] D. M. Kreps and R. B. Wilson. Sequential equilibria. *Econometrica*, 50:863–894, 1982.

[11] M. L. Littman and P. Stone. A polynomial-time Nash equilibrium algorithm for repeated games. *Decision Support Systems*, 39(1):55–66, 2005.

[12] M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[13] A. Rubinstein. Equilibrium in supergames with the overtaking criterion. *Journal of Economic Theory*, 21(1):1–9, 1979.

[14] R. Selten. Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit. *Zeitschrift für Gesamte Staatswissenschaft*, 121:301–324 and 667–689, 1965.