On the Outcome Equivalence of Extensive-Form and Behavioral Correlated Equilibria

Brian Hu Zhang[1], **Tuomas Sandholm**[1,2,3,4]
[1]Computer Science Department, Carnegie Mellon University
[2]Strategy Robot, Inc.
[3]Strategic Machine, Inc.
[4]Optimized Markets, Inc.
{bhzhang, sandholm}@cs.cmu.edu

Abstract

We investigate two notions of correlated equilibrium for extensive-form games: extensive-form correlated equilibrium (EFCE) and behavioral correlated equilibrium (BCE). We show that the two are outcome-equivalent, in the sense that every outcome distribution achievable under one notion is achievable under the other. Our result implies, to our knowledge, the first polynomial-time algorithm for computing a BCE.

1 Introduction

Computing a Nash equilibrium is hard in general-sum games, even for normal-form games with two players (Chen, Deng, and Teng 2009). Further, Nash equilibrium assumes that the players are playing *independently*, which may not hold in practice—players may have access to a shared source of randomness, or to a mediator that allows them to correlate their behavior. These concerns motivate the definition of notions of *correlation* in games.

A *normal-form correlated equilibrium* (NFCE) (Aumann 1974) is a distribution over strategy profiles from which a player, after receiving a recommended strategy from this distribution, has no incentive to disobey that recommendation. This notion, although reasonable in normal-form games, is unsuitable for extensive-form games, for at least two reasons: first, no polynomial-time algorithm is known for computing a normal-form correlated equilibrium in an extensive-form game; second, a player seeking a profitable deviation from an NFCE can condition its play on the *entire game strategy* recommended by the mediator. In large games, this is not only computationally difficult but also hard to justify. Therefore, several notions of correlation have emerged for extensive-form games, as reasonable generalizations of the NFCE to extensive-form games.

In this paper, we focus on two such notions: the *extensive-form correlated equilibrium* (EFCE) (von Stengel and Forges 2008) and the *behavioral correlated equilibrium* (BCE)[1] (Morrill et al. 2021b; Zhang 2022). In both notions, a strategy profile is first sampled from a known distribution. A player, upon reaching any of its decision points, observes only the *local* recommendation given by the strategy profile at that decision point. The distribution is considered an equilibrium if no player has incentive to disobey any recommendations.

The two notions differ in how they treat players who have disobeyed a past recommendation. In EFCE, a player who deviates from a recommendation receives no further recommendations from the mediator. In BCE, a player who disobeys a recommendation *continues to receive recommendations, and must be incentivized to follow those recommendations even though it has deviated in the past*. These conditions would seemingly make BCE a stronger notion than EFCE: a deviating player both gets more information (in the form of extra recommendations) and has a stronger incentive constraint (they must be incentivized to obey the extra recommendations). Indeed, Morrill et al. (2021b) show an explicit example (which we discuss in Section 3) of a BCE that is not an EFCE.

There are several known techniques for computing EFCEs and BCEs. Jiang and Leyton-Brown (2011) developed an exact polynomial-time algorithm that finds an EFCE. Celli et al. (2020) developed polynomial-time no-regret dynamics that converge to EFCE at rate $\mathrm{poly}(|H|, 1/\varepsilon)$ where $|H|$ is the number of nodes in the game tree. The main technique for computing BCE is no-regret learning. Morrill et al. (2021b) and Zhang (2022) independently developed very similar algorithms for computing BCEs via no-regret learning. Both of their algorithms, however, take time $\mathrm{poly}(b^d, 1/\varepsilon)$ where b is the branching factor and d is the depth—when $|H| \ll b^d$, this is exponential in the size of the game[2]. The discrepancy between these bounds has led Song, Mei, and Bai (2022) to define the K-EFCE, which interpolates between EFCE ($K = 1$) and BCE ($K = d$) by allowing a deviating player K deviations before it stops receiving recomendations. They develop no-regret learning dynamics with convergence rate $\mathrm{poly}(|H|, b^K, 1/\varepsilon)$, thus matching the known results for EFCE and BCE. To our knowledge, finding a BCE in time $\mathrm{poly}(|H|, 1/\varepsilon)$ was an open problem.

Our main result is that, at least in some sense, the distinctions between EFCE and BCE are insignificant. More formally, we show that *every EFCE can be transformed into*

[1]This notion was independently defined by the two papers cited; Zhang (2022) uses the name *forgiving correlated equilibrium*.

[2]Zhang (2022) states their algorithm as having polynomial runtime, because their paper assumes uniform depth and branching factor (so that $|H| = \Theta(b^d)$).

a BCE that achieves the same outcome distribution—that is, the same distribution over terminal nodes—and moreover that there is a polynomial-time algorithm for implementing such a transformation. Our result implies, to our knowledge, the first polynomial-time algorithm for computing a BCE in an extensive-form game.

2 Preliminaries

We now introduce the notions necessary for this paper.

Extensive-Form Games

An n-player *extensive-form game* consists of the following.

1. a tree of histories H, rooted at $\varnothing$. The set of leaves, or *terminal histories*, is denoted Z. The edges of H are labeled with *actions*, and for a node $h \in H \setminus Z$, the set of actions at h is denoted A_h;
2. a partition $H \setminus Z = H_0 \sqcup H_1 \sqcup \cdots \sqcup H_n$ of the histories, where H_i is the set of nodes at which player i acts;
3. for each player $i \in [n] := \{1, \ldots, n\}$, a partition $\mathcal{I}_i$ of H_i into *information sets*, or *infosets*. Nodes in the same information set must have the same set of action labels: for an information set $I \in \mathcal{I}_i$, the shared action set is denoted A_I;
4. for each node $h \in H_0$, a fixed distribution $p(\cdot|h)$ over the actions at h; and
5. for each player i, a *utility function* $u_i : Z \to \mathbb{R}$.

We will demand that players have *perfect recall*, in other words, that they do not forget information. Formally, call $\sigma_i(h)$ the *sequence* of information sets reached by player i and actions played at those infosets, on the path from the root to node h, not including (if any) the infoset at h itself. We use Σ_i to denote the set of all sequences of player i. Then we will insist that all nodes in the same infoset $I \in \mathcal{I}_i$ have the same sequence for player i, and we will write $\sigma_i(I)$ to denote that shared sequence. In perfect-recall games, the last infoset-action pair uniquely identifies a sequence; therefore, we will write Ia to mean the sequence ending with the infoset I and action a.

The game tree induces a natural partial ordering over infosets, sequences, and histories. We will use $\preceq$ to denote this ordering. For example, $I \preceq z$ means z is a descendant of some $h \in I$, and $Ia' \preceq I'$.

A *pure strategy* $x_i \in X_i$ assigns an action $a \in A_I$ to each infoset $I \in \mathcal{I}_i$. A *pure strategy profile* (or simply *pure profile*) $x = (x_1, \ldots, x_n)$ is a tuple of pure strategies, one per player. $-i$ denotes the set of all players except i—for example, $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. Notationally, we will write $x_i(a|I) = 1$ if strategy x_i plays action a at infoset I (and 0 otherwise). Analogously, we will write $x_i(t|s) = 1$ if player i plays all the actions on the path from s to t (both s and t could be nodes, infosets, or sequences), and $x_i(s) = 1$ if $x_i(s|\varnothing) = 1$. Note that $x_i(Ia)$ and $x_i(a|I)$ are different: the former is the indicator that *sequence* Ia is reached by player i, whereas the latter is the indicator of whether action a is played *locally* at infoset I. We will also write $x(z) := \prod_{i \in [n]} x_i(z)$ and $x_{-i}(z) := \prod_{j \neq i} x_j(z)$.

A *mixed strategy* of player i is a distribution $\pi_i \in \Delta(X_i)$.[3] We say π_i is *behavioral* if the actions at every infoset of player i are mutually independent.

A *correlated strategy profile* $\pi \in \Delta(X_1 \times \cdots \times X_n)$ is a distribution over pure strategy profiles. Any correlated strategy profile induces a distribution over the terminal nodes of the game. We will call this distribution the *outcome distribution* induced by π, and we use $z \sim \pi$ (or $z \sim x$ if $\pi = x$ happens to be a pure profile) to denote a sample from it.

Given any pure strategy profile x, the (expected) utility $u_i(x)$ of player i is

$$u_i(x) := \mathop{\mathbb{E}}_{z \sim x} u_i(z) = \sum_{z \in Z} u_i(z) p(z) x(z)$$

where $p(z)$ is the probability that chance plays all actions on the $\varnothing \to z$ path. It will also be useful to define the *counterfactual utility*. Intuitively, $u_i(x; I)$ is the conditional utility that player i achieves at infoset I, multiplied by the probability that *other players* reach infoset I. Given a player i and infoset $I \in \mathcal{I}_i$, the counterfactual utility from I is defined by

$$u_i(x; I) := \sum_{z \succ I} u_i(z) x_i(z|I) x_{-i}(z).$$

To avoid issues of bit complexity, we assume that all numbers (utilities, nature reach probabilities, correlated strategy profiles, *etc.*) are expressed as rational numbers with $\mathrm{poly}(|H|)$-bit numerators and denominators.

Extensive-Form and Behavioral Correlated Equilibria

To define the notions of equilibrium relevant to this paper, we must first introduce the framework of Φ-regret (Greenwald and Jafari 2003). For each player i, let Φ_i^* be the set of functions $\phi : X_i \to X_i$. A function $\phi \in \Phi_i^*$ should be interpreted as a *deviation* by player i: if player i should play x under π, it instead will play $\phi(x)$.

Definition 2.1. Let π be a correlated profile. The *regret* of player i against $\phi : X_i \to X_i$ is the amount by which i would increase its expected utility by applying deviation ϕ:

$$R_i(\pi, \phi) := \mathop{\mathbb{E}}_{x \sim \pi} [u_i(\phi(x_i), x_{-i}) - u_i(x_i, x_{-i})].$$

For each player i, let $\Phi_i \subseteq \Phi_i^*$ be a set of deviations. Let $\Phi = (\Phi_1, \ldots, \Phi_n)$. We say that π is an (ε, Φ)–*equilibrium* if no deviation in Φ is more than ε-profitable, that is, if $R_i(\pi, \phi) \leq \varepsilon$ for every player i and deviation $\phi \in \Phi_i$.

Larger sets Φ_i create tighter notions of equilibrium. For example, if each Φ_i is the set of constant transformations, $\Phi_i = \{\phi : x \mapsto x^* \mid x^* \in X_i\}$,[4] then a Φ-equilibrium is a *normal-form coarse correlated equilibrium* (Moulin and Vial 1978); if $\Phi_i = \Phi_i^*$ for every i, then a Φ-equilibrium is a *normal-form correlated equilibrium* (Aumann 1974). The notions of interest to us in this paper will lie between these two extremes.

We may also enforce the above condition *from any infoset*, leading to a notion of *counterfactual* Φ*-regret*.

[3] $\Delta(S)$ is the set of probability distributions on S.

[4] $\phi : x \mapsto x^*$ is the function that maps every input to x^*.

Definition 2.2 (Morrill et al. 2021a)**.** The *counterfactual regret* of player i against deviation $\phi : X_i \to X_i$ at infoset I is the amount by which player i would increase its counterfactual utility from I by applying ϕ:

$$R_i(\pi, \phi; I) := \mathop{\mathbb{E}}_{x \sim \pi}[u_i(\phi(x_i), x_{-i}; I) - u_i(x_i, x_{-i}; I)]$$

A *counterfactual* (ε, Φ)*-equilibrium* is a correlated profile π such that no deviation in Φ has more than ε counterfactual regret from any infosets, that is, if $R_i(\pi, \phi; I) \leq \varepsilon$ for every $i \in [n]$, $\phi \in \Phi_i$, and $I \in \mathcal{I}_i$.

We now define two relevant sets of deviations, one of which uses the usual (non-counterfactual) notion of regret, and the other of which uses the counterfactual regret.

Definition 2.3 (von Stengel and Forges 2008; Morrill et al. 2021b)**.** A *causal deviation* is a deviation that can be executed by a player who, upon reaching any infoset I, observes the recommendation $x_i(\cdot|I)$ before choosing its action, *unless it has disobeyed a past recommendation*. More formally, a causal deviation is a function $\phi \in \Phi_i^*$ such that $\phi(x_i)(\cdot|I)$ depends only on I, and the values $x_i(Ja)$ for $J \preceq I$. An *ε-extensive-form correlated equilibrium* (EFCE) is an (ε, Φ)-equilibrium where Φ is the set of causal deviations.

The extensive-form correlated equilibrium is a well-understood notion. It is known that the complexity of computing one EFCE exactly is polynomial (Jiang and Leyton-Brown 2011), and more recently, regret minimization algorithms have emerged that are guaranteed to converge to EFCE (Celli et al. 2020).

Definition 2.4 (Morrill et al. 2021b)**.** A *behavioral deviation*[5] is a deviation that can be executed by a player who, upon reaching any of its infosets I, observes the recommendation $x_i(\cdot|I)$ before choosing its action. More formally, a behavioral deviation is a function $\phi \in \Phi_i^*$ such that $\phi(x_i)(\cdot|I)$ depends only on I, and the values $x_i(\cdot|J)$ for $J \preceq I$. An *ε-behavioral correlated equilibrium* (BCE) is a counterfactual (ε, Φ)-equilibrium where Φ is the set of behavioral deviations.

It is clear from the definitions that every BCE is an EFCE. BCE appears at first to be a significant refinement of EFCE. Indeed, the definition refines EFCE in two ways. First, BCE uses a larger family of deviations (every causal deviation is behavioral, but not the other way); second, BCE uses *counterfactual* regret whereas EFCE uses only the typical Φ-regret. Indeed, three disjoint sets of authors (Morrill et al. 2021b; Song, Mei, and Bai 2022; Zhang 2022) have developed no-regret learning algorithms converging to behavioral correlated equilibrium. However, unlike the aforementioned EFCE algorithms, these algorithms have worst-case runtime exponential in the size of Γ.

3 Main Result Statement and Examples

We start by defining our notion of outcome equivalence:

Definition 3.1. Two correlated strategy profiles π and π' are *outcome-equivalent* if they induce the same outcome distribution.

[5]One should not confuse behavioral *deviations* from behavioral *strategies*—the two terms only share a name.

Our main result, then, simply states:

Theorem 3.2 (Main result)**.** *Every ε-EFCE is outcome-equivalent to an ε-BCE.*

Before proving the main result, we give two examples showing why such a result may be believable and illustrating some of the ideas used in the proof. The first, due to Morrill et al. (2021a) gives an example of a BCE that is not an EFCE.

Example 3.3 (Morrill et al. 2021a). Consider the extensive-form game depicted in Figure 1 (left). Consider the correlated profile π that uniformly mixes between the profiles $(\neg U, X_1|U, X_1|\neg U, X_2)$ and $(\neg U, Y_1|U, Y_1|\neg U, Y_2)$. Both players achieve expected utility 1.5. This profile is *not* a BCE: player 1 can deviate profitably by playing U (contrary to the recommendation), and then following the recommendation to play either X_1 or Y_1. However, this deviation does not work for EFCE, because a player who deviates by playing U will not receive the subsequent recommendation. Indeed, π *is* an EFCE. This shows that behavioral deviations can be more powerful than causal deviations.

Although π is not a BCE, there is a BCE π' that is *outcome-equivalent* to π. Indeed, consider the correlated profile that evenly mixes between $(\neg U, X_1|U, X_1|\neg U, X_2)$ and $(\neg U, X_1|U, Y_1|\neg U, Y_2)$ (where the only difference is that, in the second pure profile, $Y_1|U$ has been replaced by $X_1|U$). This change preserves the outcome distribution, because the recommendation $Y_1|U$ is never actually given to player 1 in equilibrium, as player 1 plays $\neg U$ in equilibrium. This profile π' *is* a BCE: the previous deviation no longer works, because, after playing U, player 1 is always given the recommendation X_1—its counterfactual best response—instead of any useful recommendation.

The second example shows that the use of *counterfactual* regret in the BCE definition is also significant.

Example 3.4. Consider the (one-player) extensive-form game depicted in Figure 1 (right). Then the profile $\pi = (0.9L + 0.1R, R')$ is a 0.2-EFCE, but it is not an 0.2-BCE, because the player can deviate to L' at B to counterfactually improve its utility at that infoset by 1. However, once again, there is a 0.2-BCE that is outcome-equivalent to π: namely, $\pi' = 0.9(L, L') + 0.1(R, R')$.

Interestingly, in this example, the profile π' is not a behavior strategy, and indeed there is no 0.2-BCE that is a behavior strategy and outcome-equivalent to π. This illustrates that converting from EFCE to BCE in general will sometimes require turning behavior strategies into non-behavior strategies.

4 Proof of Main Result

In this section, we prove the main result, Theorem 3.2.

Let π be an ε-EFCE. For each player i, infoset $I \in \mathcal{I}_i$ and action $a \in A_I$, define the *counterfactual best response strategy* x_i^{Ia} as the strategy that maximizes the countefactual utility at I against π_{-i}, conditioned on x_i playing to Ia. Formally,

$$x_i^{Ia} \in \mathop{\text{argmax}}_{x_i' \in X_i} \mathop{\mathbb{E}}_{x \sim \pi}[u_i(x_i', x_{-i}; I) \mid x_i(Ia) = 1].$$

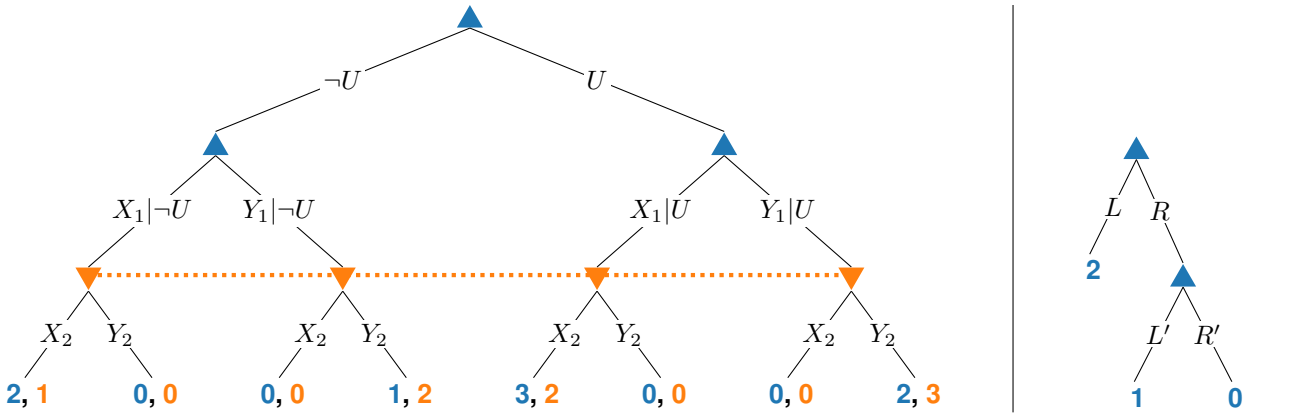

Figure 1: *Left:* The *extended battle of the sexes* game in Example 3.3. The players are ▲ (P1) and ▼ (P2). Infosets are connected by dotted lines. Player 1 first chooses whether or not to *upgrade* (U). Then, both players simultaneously choose an event (X or Y) to attend. Player 1 prefers X, while player 2 prefers Y. If the players attend different events, they are unhappy and get utility 0. If the players attend the same event, the player attending its preferred event gets 2, and the other player gets 1. Upgrading gives an extra point of utility if the players attend the same event. *Right:* The game used in Example 3.4 illustrating that use counterfactual regret is also significant.

for every infoset I. Assume ties are broken consistently—for example, in favor of the lexicographically first action. Of course, it is only interesting to investigate x_i^{Ia} at infosets $J \succeq I$. Given the conditional opponent reach probabilities

$$\mathop{\mathbb{E}}_{x\sim\pi}[x_{-i}(z)|x_i(Ia)=1]$$

for every $z \in Z$, the strategy x_i^{Ia} can be computed by a simple backwards tree traversal.

Now, consider the distribution π' generated by the following procedure. Sample $x \sim \pi$, and then for every infoset $I \in \mathcal{I}_i$ not reached, replace the recommendation at I with the recommendation at I in x_i^{Ja} where player i deviated before I. Formally, for every player i and every infoset $I \in \mathcal{I}_i$ with $x_i(I) = 0$, let Ja be the sequence that i deviates before reaching I, that is, let Ja be such that $x_i(Ja) = 1$, $J \preceq I$, but $Ja \not\preceq I$. Then replace $x_i(\cdot|I)$ with $x_i^{Ja}(\cdot|I)$.

We claim that π' is an ε-BCE. Let ϕ be any behavioral deviation of player i and let I be any infoset of player i. Let $x \sim \pi'$. First, note that, by construction of π', a deviating player in π' cannot profit from behavioral deviations compared to causal deviations. This is because, for any sequence Ia, the values $x_i(Ja')$ for $J \preceq I$ completely determine $x_i(Ia)$: if $x_i(I) = 1$ then $x_i(Ia) = x_i(a|I)$, and if $x_i(I) = 0$ then $x_i(Ia) = x_i^{Ja'}(I)$ where Ja is the deviation point of x_i before I. Thus, we may assume that ϕ is causal. Further, ϕ cannot profit on $x_i \sim \pi'$ such that $x_i(I) = 0$: by definition, x_i will be playing a counterfactual best response at every such infoset I conditioned on all information that the deviating player knows at that point. In symbols,

$$\begin{aligned}
&R_i(\pi', \phi; I)\\
&= \underbrace{\mathop{\mathbb{E}}_{x\sim\pi'}[u_i(\phi(x_i), x_{-i}; I) - u_i(x_i, x_{-i}; I) \mid x(I) = 0]}_{\leq 0}\\
&\qquad \cdot \mathop{\mathbb{E}}_{x\sim\pi'}[1 - x(I)]\\
&\quad + \mathop{\mathbb{E}}_{x\sim\pi'}[u_i(\phi(x_i), x_{-i}; I) - u_i(x_i, x_{-i}; I) \mid x(I) = 1]\\
&\qquad \cdot \mathop{\mathbb{E}}_{x\sim\pi'}[x(I)]\\
&\leq \mathop{\mathbb{E}}_{x\sim\pi'}[u_i(\phi(x_i), x_{-i}; I) - u_i(x_i, x_{-i}; I) \mid x(I) = 1]\\
&\qquad \cdot \mathop{\mathbb{E}}_{x\sim\pi'}[x(I)]\\
&= R_i(\pi', \phi^{\succeq I}) \leq \varepsilon
\end{aligned}$$

where $\phi^{\succeq I}$ is the deviation that applies ϕ only at infosets $J \succeq I$, that is,

$$\phi^{\succeq I}(x_i)(a|J) = \begin{cases} \phi(x_i)(a|J) & \text{if} \quad J \succeq I \\ x_i(a|J) & \text{otherwise} \end{cases}$$

□

5 Algorithm for Converting from EFCE to BCE

The proof of Theorem 3.2 also implies a *polynomial-time algorithm* for computing a BCE from an EFCE. That is, so long as the EFCE π is expressed in a form allowing for the computation of the counterfactual best responses x_i^{Ia}, the proof gives a polynomial-time algorithm for computing a BCE. In this section, we give a possible formulation of this polynomial-time algorithm. First, we must define the format that we will use to represent correlated profiles.

Definition 5.1. A correlated profile π is a *mixture of small-*

support products if

$$\pi = \sum_{t=1}^{T} \alpha^{(t)} \bigotimes_{i=1}^{n} \pi_i^{(t)} \quad \text{where} \quad \pi_i^{(t)} = \sum_{k=1}^{K} \beta_i^{(t,k)} x_i^{(t,k)}.$$

where T and K are positive integers, $\sum_{t=1}^{T} \alpha^{(t)} = 1$, $\sum_{k=1}^{K} \beta_i^{(t,k)} = 1$ for every i and t, and the notation $\bigotimes_{i=1}^{n} \pi_i$ means the product distribution $\pi \in \Delta(X_1) \times \cdots \times \Delta(X_n)$ whose marginal on X_i is π_i.

Such a π can be expressed using $O(T \cdot K \cdot |H|)$ numbers, namely, for each $t \le T, k \le K$, and sequence $Ia \in \Sigma_i$ we need to represent $x_i^{(t,k)}(a|I)$, $\alpha^{(t)}$, and $\beta_i^{(t,k)}$. We will assume in the rest of this section that correlated profiles are represented as a mixture of products.

One may wonder at this point about the case where the $\pi_i^{(t)}$ are behavior strategies. In this case, it is possible for K to be exponentially large: for example, if $\pi_i^{(t)}$ is fully mixed then $K = \prod_{I \in \mathcal{I}_i} |A_I|$. However, there is a remedy for this:

Lemma 5.2. *Let π be an ε-EFCE expressed as a mixture of T products, where each $\pi_i^{(t)}$ is a behavior strategy. Then, there is a $\mathrm{poly}(|H|, T)$-time algorithm that returns an ε-EFCE π' that (1) is outcome-equivalent to π, and (2) is a mixture of small-support products with $K \le |H|$ and the same T.*

Proof. By definition, the EFCE gap and the outcome distribution both only depend on the *sequence-form reach probabilities* $\pi_i^{(t)}(Ia)$ for each $Ia \in \Sigma_i$. These form a vector $\boldsymbol{\pi}_i^{(t)} \in [0,1]^{\Sigma_i}$, called the *sequence-form vector*. Intuitively, the sequence-form vector $\boldsymbol{\pi}_i^{(t)}$ is a complete description of a strategy up to outcome equivalence, since the probability of any given terminal node being reached under profile π is just $p(z) \cdot \prod_i \boldsymbol{\pi}_i^{(t)}(z)$. Therefore, it suffices to show that the sequence-form vector $\boldsymbol{\pi}_i^{(t)}$ is a convex combination of a small number of sequence-form vectors of pure strategies. The fact that the set of sequence-form vectors is a convex polytope in extensive-form games was shown by Koller, Megiddo, and von Stengel (1994). Indeed, one can directly describe the set using the following linear constraint system:

$$\forall \sigma\colon \pi_i(\sigma) \ge 0; \quad \pi_i(\varnothing) = 1; \quad \forall I\colon \pi_i(I) = \sum_{a \in A_I} \pi_i(Ia).$$

Therefore, by Carathéodory's theorem[6] on convex hulls, there exists a decomposition $\boldsymbol{\pi}_i^{(t)} = \sum_{k=1}^{|\Sigma_i|} \beta_i^{(t,k)} \boldsymbol{x}_i^{(t,k)}$, where the $\boldsymbol{x}_i$s are sequence forms of pure strategies. An explicit algorithm for computing such a decomposition is described by Grötschel, Lovász, and Schrijver (1981, Theorem 3.9). This completes the proof. □

All algorithms that we are aware of that compute an exact or approximate EFCE return their correlated profiles as mixtures of behavioral profiles (or as mixtures of pure profiles, which are just the special case $K = 1$). Lemma 5.2 is therefore important in that it allows us to convert from behavior strategies to mixtures of small-support products, and therefore allows the main result of this section to also deal with behavior strategy profiles. We are now ready to state the main result of this section.

Theorem 5.3. *There exists a $\mathrm{poly}(|H|, T, K)$-time algorithm that takes as input an ε-EFCE π as a mixture of small-support products, and returns ε-BCE in the same format.*

Proof. Follow the proof of Theorem 3.2. The counterfactual best responses x_i^{Ia} can be computed in polynomial time because one can compute $\mathbb{E}_{x \sim \pi}[x_{-i}(z) \mid x_i(Ia) = 1]$ for every $z \in Z$ by iterating over the support of π. Then, for each $x_i^{(t,k)}$, replacing $x_i^{(t,k)}(\cdot|I)$ with $x_i^{Ja}(\cdot|I)$ as directed by Theorem 3.2 is a matter of iterating over the information sets of player i and keeping track of where deviations happen. □

In particular, applying the polynomial-time exact EFCE algorithm of Jiang and Leyton-Brown (2011), we have:

Corollary 5.4. *There is a polynomial-time algorithm that, given an extensive-form game, computes an exact BCE.*

To our knowledge, this result was previously unknown, even for ε-approximate BCE, not to mention exact BCE.

6 Discussion

In this section, we discuss a corollary and some caveats to our results and techniques.

Optimal Equilibria

Our results have immediate implications for the problem of *optimizing* over the set of BCEs. Let $c : Z \to \mathbb{R}$ be any objective function. Call an equilibrium π *optimal* under objective c if it maximizes $c(\pi) := \mathbb{E}_{x \sim \pi, z \sim x}\, c(z)$ among all equilibria of the same notion. The following corollary follows immediately from Theorem 3.2.

Corollary 6.1. *For every objective c, the optimal EFCE and the optimal BCE under c have the same objective value.*

Therefore, to compute an optimal BCE, it suffices to compute an optimal EFCE and convert it to a BCE. The conversion can be perfomed in polynomial time by Theorem 5.3. As for computing an optimal EFCE, the general problem is NP-hard (von Stengel and Forges 2008), but various algorithms exist for the optimal EFCE problem that have parameterized guarantees (Zhang et al. 2022) or work in special cases (Farina and Sandholm 2020). Our results therefore imply, up to polynomial factors, algorithms with the same guarantees for optimal BCE.

Hindsight Rationality and No-Regret Learning

So far, this paper has only discussed algorithms that take an already-computed ε-EFCE *as input*. However, one possible motivation of notions of correlated equilibria is that uncoupled learning dynamics can reach them in empirical frequency of play. Formally, suppose that n agents play an extensive-form game repeatedly for T rounds. At time $t \in [T]$, each agent i selects a (usually behavior) strategy $\pi_i^{(t)} \in \Delta(X_i)$, which depends only on the players' *own*

[6] Carathéodory's theorem states that every point in a convex compact set X of dimension d is a convex combination of at most $d+1$ extreme points of X.

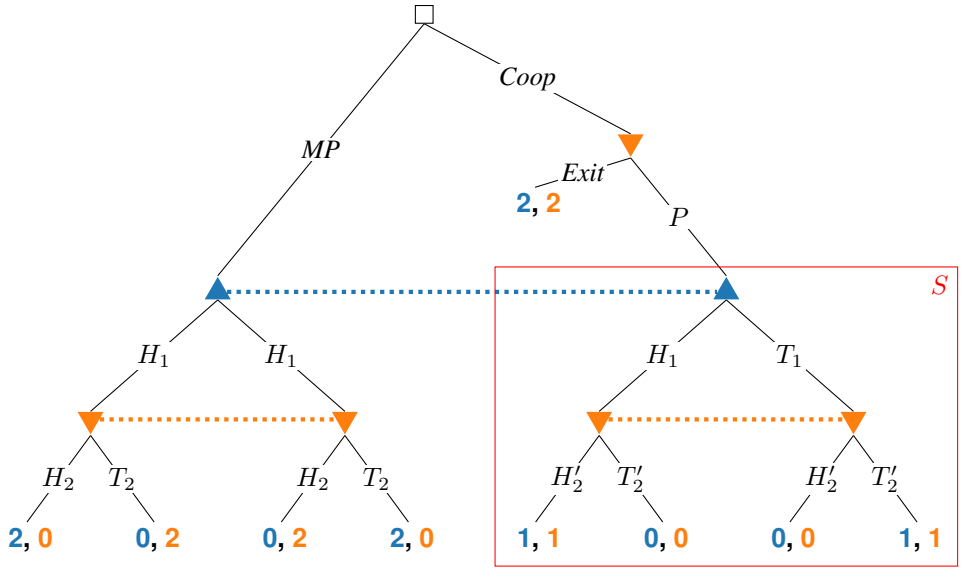

Figure 2: A game showing that the EFCE-BCE map in this paper is not surjective (for any tiebreaking method). The root node is a nature node; nature moves uniformly at random. The *MP* subtree is the matching pennies game; the *Coord* subtree is a coordination game, but P2 has a strictly dominant *Exit* action.

utility functions $u_i(\cdot, \pi_{-i}^{(\tau)})$ for each $\tau < t$. (in particular, not on the other players' utility functions). Then the *empirical frequency of play* is the uniform distribution on $\{\pi^{(1)}, \dots, \pi^{(T)}\}$.

As stated earlier, there are known uncoupled learning dynamics that approach an ε-EFCE after $T = \text{poly}(|H|, 1/\varepsilon)$ rounds (Celli et al. 2020). However, to our knowledge, there is no known learning algorithm whose empirical frequency of play approaches BCE at $\text{poly}(|H|, 1/\varepsilon)$: the earlier algorithms of Morrill et al. (2021b); Song, Mei, and Bai (2022); and Zhang (2022) achieve rate $\text{poly}(b^d, 1/\varepsilon)$, where b is the depth and d is the branching factor of the game, but this is worst-case exponential in the size of the game. Roughly speaking, the reason for the discrepancy is that a deviator seeking a profitable BCE deviation has $\text{poly}(b^d)$ possible decision points (corresponding to each sequence of recommendations it could have seen), whereas a deviator seeking a profitable EFCE deviation only has polynomially many (because, at each infoset I, the only possible recommendation histories are the sequences Ja for $J \preceq I$).

One may therefore ask whether Theorem 5.3 implies the existence of uncoupled learning dynamics that reach BCE in $\text{poly}(|H|, 1/\varepsilon)$ rounds. Unfortunately, this is not the case. Theorem 3.2 (and therefore the algorithm in Theorem 5.3) changes player i's strategy $\pi_i^{(t)}$ based on *future* strategies $\pi_{-i}^{(>t)}$, because the counterfactual best response x_i^{Ia} depends on *all* opponent strategy profiles, not just those played in the past. We leave finding polynomial-time uncoupled learning dynamics for BCE (or proving the nonexistence of such dynamics) as an open question for future research.

Stronger Notions of Outcome Equivalence

The notion of outcome equivalence used throughout the paper so far concerns only the outcome distribution *on the equilibrium path of play*. One may ask whether this notion can be strengthened, and what happens to our results under a stricter definition of outcome equivalence. For example, one may consider the following strengthened notion: let us call two profiles π and π' *counterfactually outcome-equivalent* if, for every player i and infoset $I \in \mathcal{I}_i$, the counterfactual reach probabilities of every terminal node $z \succ I$ coincide, that is,

$$\mathop{\mathbb{E}}_{x \sim \pi} x_i(z|I) x_{-i}(z) = \mathop{\mathbb{E}}_{x \sim \pi'} x_i(z|I) x_{-i}(z).$$

This would guarantee, among other things, that the counterfactual utility $\mathbb{E}\, u_i(x; I)$ is the same under π and π' at every infoset. Unfortunately but perhaps unsurprisingly, our main result does not hold for this stronger notion of outcome equivalence. Indeed, consider the same game used in Example 3.4 (Figure 1, right). In that game, there exists an EFCE, namely the pure strategy (L, R'), whose counterfactual value at the lower P1 decision point is zero. There cannot be a BCE with this property, because then there would be a beneficial counterfactual deviation at that decision point.

We define the above notion of counterfactual outcome-equivalence purely for simplicity, as we have been using counterfactual utility throughout the paper. However, the above counterexample would also apply equally well to other possible strengthenings of the notion of equivalence, such as a subgame-perfect notion (*e.g.*, "the conditional distributions coincide in every proper subgame").

Surjectivity

Let f be the map in Theorem 3.2, that is, f takes as input an ε-EFCE π and outputs an outcome-equivalent ε-BCE $f(\pi)$.

Every BCE *outcome distribution* appears as the outcome distribution of some $f(\pi)$: f preserves outcome distributions, so taking π to be a BCE with the desired outcome distribution is sufficient. However, one may ask whether the map given by Theorem 3.2 is surjective on the set of all BCEs, not just the set of all outcome distributions—that is, whether every BCE appears as $f(\pi)$. The map f depends on a choice of tiebreaking scheme for the counterfactual best responses x_i^{Ia}. In this section, we give a simple counterexample illustrating that, *regardless of the tiebreaking scheme*, f cannot be surjective. Consider the game in Figure 2. There exists a BCE of this game in which P1 gets conditional utility 1 in the subtree S, namely the uniform distribution on (E, H_1, H_2, H_2'), (E, H_1, T_2, H_2'), (E, T_1, T_2, T_2'), and (E, T_1, H_2, T_2'), that is, P2 perfectly correlates with P1 in S. However, this cannot happen in a BCE created by f because such a BCE cannot contain a useful recommendation to P2 in S because P2 must have deviated before reaching S.

Counterfactual Regret and the Definition of BCE

We discuss here the choice and consequences of using the *counterfactual regret* (Definition 2.2), rather than the usual notion of regret (Definition 2.1), in the definition of BCE.

The definitions of equilibria used throughout this paper are not new to this paper. Instead, EFCE and BCE are defined by von Stengel and Forges (2008) and Morrill et al. (2021b), respectively. Compared to EFCE, BCE enforces a sort of *equilibrium refinement*—not quite subgame perfection, but something resembling it. Further, for no-regret algorithms in particular, using counterfactual regret is fairly natural—indeed, the *couterfactual regret minimization (CFR)* family of algorithms (Zinkevich et al. 2007)—as its name suggests—entirely revolves around mimimizing the counterfactual regret, and many of the best equilibrium-finding algorithms for extensive-form games are based on CFR (Brown and Sandholm 2019; Farina, Kroer, and Sandholm 2021).

One may indeed define a notion of equilibrium that is like BCE except that it uses regular regret (Definition 2.1) instead of counterfactual regret. Let us call this notion *full EFCE*. As this choice of name suggests, full EFCE behaves more like EFCE than BCE. Indeed, von Stengel and Forges (2008) originally define the full EFCE (although they do not give it a special name), and they then show that full EFCE and EFCE are outcome-equivalent, before using what we define as the EFCE for the remainder of their paper. Intuitively, the outcome equivalence follows from a conversion in which the recommendations in off-path information sets are replaced with arbitrary (uninformative) recommendations (since they are off-path, there is no need to ensure incentive compatibility). Further, the outcome equivalence between full EFCE and EFCE—unlike the one shown in our paper between BCE and EFCE—also carries over to hindsight rationality, so the polynomial-time no-regret algorithms that converge to EFCE (*e.g.*, Celli et al. 2020) can be easily modified to converge to full EFCE instead. Due to this equivalence, subsequent papers, including all those cited in our paper, have used the same definition of EFCE that we use, as it is simpler. BCE, on the other hand, has no known polynomial-time no-regret dynamics. We leave this as an explicit open problem.

7 Conclusions and Future Research

We have proven the outcome equivalence of extensive-form and behavioral correlated equilibria, and we have given a polynomial-time algorithm for converting one into the other, thus leading to, among other implications, the first algorithm for computing a BCE in polynomial time.

Perhaps the most relevant question for future research is whether there are uncoupled learning dynamics converging to BCE at rate $\mathrm{poly}(|H|, 1/\varepsilon)$. Resolving this question in either direction would be illuminating. If there are, then the algorithm would somehow overcome the exponential explosion in the number of decision points accessible to the deviator. If there are not, then BCE would be a rare example of a notion of equilibrium for which finding an equilibrium is doable in polynomial time, but not with uncoupled learning dynamics.

More broadly, there remains an interesting open line of research regarding the limits of polynomial-time algorithms for computing equilibria: how tight does one need to make the notion before computing one becomes hard? For example, computing a Nash equilibrium is known to be hard (Chen, Deng, and Teng 2009). What about the normal-form correlated equilibrium (NFCE) (Aumann 1974) in an extensive-form game, which lies between EFCE and Nash? Is there a polynomial-time algorithm for finding one? Are there polynomial-time uncoupled learning dynamics that converge to one at rate $\mathrm{poly}(|H|, 1/\varepsilon)$?[7] Do these answers change if we instead define and use a *counterfactual* notion of NFCE, which would then lie between BCE and a counterfactual notion of Nash equilibrium (in which each player's strategy must be a *counterfactual* best response to other players' strategies, that is, each player must be best responding even at infosets I that the player does not play to reach, as long as other players play to reach I)? All these questions are, to our knowledge, open.

Acknowledgements

We thank Amy Greenwald, Yu Bai, and Hugh Zhang for very helpful discussions. This material is based on work supported by the Vannevar Bush Faculty Fellowship ONR N00014-23-1-2876, National Science Foundation grants RI-2312342 and RI-1901403, ARO award W911NF2210266, and NIH award A240108S001.

References

Aumann, R. 1974. Subjectivity and Correlation in Randomized Strategies. *Journal of Mathematical Economics*, 1: 67–96.

Brown, N.; and Sandholm, T. 2019. Solving imperfect-information games via discounted regret minimization. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Celli, A.; Marchesi, A.; Farina, G.; and Gatti, N. 2020. No-regret learning dynamics for extensive-form correlated equi-

[7] A recent simultaneous breakthrough by Peng and Rubinstein (2023) and Dagan et al. (2023), which appeared after we submitted the present paper, has shown an algorithm whose convergence rate is roughly $|H|^{\tilde{O}(1/\varepsilon)}$, but the $\mathrm{poly}(|H|, 1/\varepsilon)$ case remains open.

librium. *Conference on Neural Information Processing Systems (NeurIPS)*, 33: 7722–7732.

Chen, X.; Deng, X.; and Teng, S.-H. 2009. Settling the Complexity of Computing Two-Player Nash Equilibria. *Journal of the ACM*.

Dagan, Y.; Daskalakis, C.; Fishelson, M.; and Golowich, N. 2023. From External to Swap Regret 2.0: An Efficient Reduction and Oblivious Adversary for Large Action Spaces. *arXiv preprint arXiv:2310.19786*.

Farina, G.; Kroer, C.; and Sandholm, T. 2021. Faster Game Solving via Predictive Blackwell Approachability: Connecting Regret Matching and Mirror Descent. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Farina, G.; and Sandholm, T. 2020. Polynomial-Time Computation of Optimal Correlated Equilibria in Two-Player Extensive-Form Games with Public Chance Moves and Beyond. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Greenwald, A.; and Jafari, A. 2003. A General Class of No-Regret Learning Algorithms and Game-Theoretic Equilibria. In *Conference on Learning Theory (COLT)*. Washington, D.C.

Grötschel, M.; Lovász, L.; and Schrijver, A. 1981. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1: 169–197.

Jiang, A.; and Leyton-Brown, K. 2011. Polynomial-time computation of exact correlated equilibrium in compact games. In *ACM Conference on Electronic Commerce (EC)*.

Koller, D.; Megiddo, N.; and von Stengel, B. 1994. Fast algorithms for finding randomized strategies in game trees. In *ACM Symposium on Theory of Computing (STOC)*.

Morrill, D.; D'Orazio, R.; Sarfati, R.; Lanctot, M.; Wright, J. R.; Greenwald, A. R.; and Bowling, M. 2021a. Hindsight and sequential rationality of correlated play. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, 5584–5594.

Morrill, D.; D'Orazio, R.; Lanctot, M.; Wright, J. R.; Bowling, M.; and Greenwald, A. R. 2021b. Efficient deviation types and learning for hindsight rationality in extensive-form games. In *International Conference on Machine Learning (ICML)*, 7818–7828. PMLR.

Moulin, H.; and Vial, J.-P. 1978. Strategically zero-sum games: The class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3-4): 201–221.

Peng, B.; and Rubinstein, A. 2023. Fast swap regret minimization and applications to approximate correlated equilibria. *arXiv preprint arXiv:2310.19647*.

Song, Z.; Mei, S.; and Bai, Y. 2022. Sample-efficient learning of correlated equilibria in extensive-form games. *Conference on Neural Information Processing Systems (NeurIPS)*, 35: 4099–4110.

von Stengel, B.; and Forges, F. 2008. Extensive-form correlated equilibrium: Definition and computational complexity. *Mathematics of Operations Research*, 33(4): 1002–1022.

Zhang, B. H.; Farina, G.; Celli, A.; and Sandholm, T. 2022. Optimal correlated equilibria in general-sum extensive-form games: Fixed-parameter algorithms, hardness, and two-sided column-generation. In *ACM Conference on Economics and Computation (EC)*, 1119–1120.

Zhang, H. 2022. A simple adaptive procedure converging to forgiving correlated equilibria. *arXiv preprint arXiv:2207.06548*.

Zinkevich, M.; Bowling, M.; Johanson, M.; and Piccione, C. 2007. Regret Minimization in Games with Incomplete Information. In *Conference on Neural Information Processing Systems (NeurIPS)*.