STEERING NO-REGRET LEARNERS TO OPTIMAL EQUILIBRIA

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the problem of steering no-regret-learning agents to play desirable equilibria via nonnegative payments. We first show that steering is impossible if the total budget (across all iterations) is finite, both in normal- and extensive-form games. However, we establish that *vanishing* average payments are compatible with steering. In particular, when players' full strategies are observed at each timestep, we show that constant per-iteration payments permit steering. In the more challenging setting where only trajectories through the game tree are observable, we show that steering is impossible with constant per-iteration payments in general extensive-form games, but possible in normal-form games or if the maximum per-iteration payment may grow with time. We supplement our theoretical positive results with experiments highlighting the efficacy of steering in large games, and show how our framework relates to optimal mechanism design and information design.

1 INTRODUCTION

Any student of game theory learns that games can have multiple *equilibria* of different quality—for example, in terms of social welfare (Figure 1). How can a *mediator*—a benevolent third party—*steer* players toward an optimal one? In this paper, we consider the problem of using a mediator who can dispense nonnegative payments and offer advice to players so as to guide to a better collective outcome.

Importantly, our theory does not rest upon strong assumptions regarding agent obedience; instead, we only assume that players have *sublinear regret*, a mild assumption on the rationality of the players adopted in several prior studies (*e.g.*, Nekipelov et al., 2015; Kolumbus & Nisan, 2022b; Camara et al., 2020). Variants of this problem have received tremendous interest in the literature (*e.g.*, Monderer & Tennenholtz, 2004; Anshelevich et al., 2008; Schulz & Moses, 2003; Agussurja & Lau, 2009; Balcan, 2011; Balcan et al., 2013; 2014; Mguni et al., 2019; Li et al., 2020; Kempe et al., 2020; Liu et al., 2022 and references therein), but prior work either operates in more restricted classes of games or makes strong assumptions regarding player obedience. We study the steering problem in its full generality for general (imperfect-information) extensive-form games under an entire hierarchy of equilibrium concepts, and we establish a number of positive algorithmic results and complementing information-theoretic impossibilities.

Summary of Our Results Our formulation enables the mediator to 1) reward players with nonnegative *payments* and 2) offer advice. Of course, with no constraints on the payments, the problem becomes trivial: the mediator could enforce any arbitrary outcome by paying players to play that outcome. On the other extreme, we show that if the *total* realized payments are constrained to be bounded, the decentralized steering problem is information-theoretically impossible (Proposition 3.2). Therefore, we compromise by allowing the total realized payments to be unbounded, but insist that the average payment per round is *vanishing*. Further, to justify 2) above, we show that *without advice*, steering to *mixed-Nash equilibria* is impossible already in normal-form games (Appendix D), although advice is not necessary for *pure-Nash* equilibria (Sections 4 and 5). Offering recommendations is in line with much of prior work (Appendix A), and is especially natural for correlated equilibrium concepts.

The goal of the mediator is to reach an equilibrium, either explicitly provided or provided as a principal utility function. We first assume that the mediator is provided an equilibrium. We distinguish between *realized* payments and *potential* payments. Realized payments are the payments actually dispensed

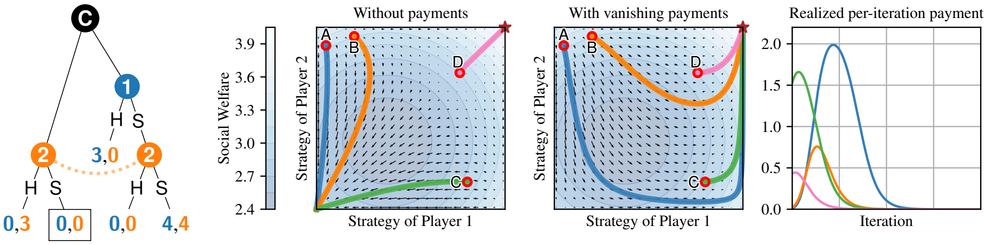


Figure 1: **Left:** An extensive-form version of a stag hunt. Chance plays uniformly at random at the root note, and the dotted line connecting the two nodes of Player 2 indicates an infoset: Player 2 cannot distinguish the two nodes. Introducing *vanishing* realized payments alters the gradient landscape, steering players to the optimal equilibrium (star) instead of the suboptimal one (opposite corner). The capital letters show the players' initial strategies. Lighter color indicates higher welfare and the star shows the highest-welfare equilibrium. Further details are in Appendix C.

to the players. *Potential* payments are payments that players *would have* received, had they played different strategies.

We first consider the *full-feedback* (Section 5) setting where players' payments may depend on players' full strategies. We present steering algorithms that establish under different computational assumptions the first main result.

Theorem (Informal; precise versions in Theorem 5.2)**.** *For both normal-form and extensive-form games, the decentralized steering problem can be solved under full feedback.*

Intuitively, the mediator sends payments in such a way as to 1) reward the player a small amount for playing the equilibrium, and 2) *compensate* the player for deviations of other players. Next, we consider the more challenging *bandit setting*, wherein only game trajectories are observed. In extensive-form games, this condition significantly restricts the structure of the payment functions, and in particular rules out the full-feedback algorithm above. We show that the decentralized steering problem under bandit feedback is information-theoretically impossible in the general case with bounded potential payments.

Theorem (Informal; precise version in Theorem 5.4)**.** *For extensive-form games, the decentralized steering problem is impossible under bandit feedback with bounded* potential *payments.*

To circumvent this lower bound, we next allow the *potential* payments to depend on the time horizon, while still insisting that they vanish in the limit.

Theorem (Informal; precise version in Theorem 5.6)**.** *For extensive-form games, if the payments may depend on the time horizon, the decentralized steering problem can be solved under bandit feedback.*

The proof of this theorem is more involved than the previous two. In particular, one might hope that the desired equilibrium can be made (strictly) dominant by adding appropriate payments as in k-implementation (Monderer & Tennenholtz, 2004). In extensive-form games, this is not the case: there are games where making the welfare-optimal equilibrium dominant would require payments in equilibrium, thereby inevitably leading to non-vanishing realized payments. Nevertheless, we show that steering is possible despite even without dominance. This leads to the intriguing behavior where some players may actually move *farther* from obedience before they move closer (compare Figure 1). As such, we significantly depart from the approach of Monderer & Tennenholtz (2004); we elaborate on this comparison and further related work in Appendix A.

Both previous positive results require computing an equilibrium upfront, which is both computationally expensive and not adaptive to players' actions. We next analyze an *online* setting, where the mediator employs an online regret minimization algorithm to compute an optimal equilibrium *while* guiding the players toward it. As expected, algorithms for the online steering problem attain slightly worse rates compared to algorithms for the offline problem. The rates we obtain for the various versions of the steering problem all decay polynomially with the number of rounds, and we highlight the time dependence in Table 1. We complement our theoretical analysis by implementing and testing our steering algorithms in several benchmark games in Section 7.

Table 1: Summary of our positive algorithmic results. We hide game-dependent constants and logarithmic factors, and assume that regret minimizers incur a (typical) average regret of $T^{-1/2}$.

	Steering to Fixed Equilibrium	Online Steering
Normal Form or Full Feedback	$T^{-1/4}$ (Theorem 5.2)	$T^{-1/6}$ (Theorem 6.5)
Extensive Form and Bandit Feedback	$T^{-1/8}$ (Theorem 5.6)	*Open problem*

2 Preliminaries

In this section, we introduce some basic background on extensive-form games.

Definition 2.1. An *extensive-form game* Γ with n players has the following components:

1. a set of players, identified with the set of integers $[\![n]\!] := \{1, \dots, n\}$. We will use $-i$, for $i \in [\![n]\!]$, to denote all players except i;
2. a directed tree H of *histories* or *nodes*, whose root is denoted $\varnothing$. The edges of H are labeled with *actions*. The set of actions legal at h is denoted A_h. Leaf nodes of H are called *terminal*, and the set of such leaves is denoted by Z;
3. a partition $H \setminus Z = H_{\mathsf{C}} \sqcup H_1 \sqcup \dots \sqcup H_n$, where H_i is the set of nodes at which i takes an action, and C denotes the chance player;
4. for each player $i \in [\![n]\!]$, a partition $\mathcal{I}_i$ of i's decision nodes H_i into *information sets*. Every node in a given information set I must have the same set of legal actions, denoted by A_I;
5. for each player i, a *utility function* $u_i : Z \to [0, 1]$ which we assume to be bounded; and
6. for each chance node $h \in H_{\mathsf{C}}$, a fixed probability distribution $c(\cdot\,|h)$ over A_h.

At a node $h \in H$, the *sequence* $\sigma_i(h)$ of an agent i is the set of all information sets encountered by agent i, and the actions played at such information sets, along the $\varnothing \to h$ path, excluding at h itself. An agent has *perfect recall* if $\sigma_i(h) = \sigma_i(h')$ for all h, h' in the same infoset. Unless otherwise stated (Section 6), we assume that all players have perfect recall. We will use $\Sigma_i := \{\sigma_i(z) : z \in Z\}$ to denote the set of all sequences of player i that correspond to terminal nodes.

A *pure strategy* of player i is a choice of one action in A_I for each information set $I \in \mathcal{I}_i$. The *sequence form* of a pure strategy is the vector $\boldsymbol{x}_i \in \{0, 1\}^{\Sigma_i}$ given by $\boldsymbol{x}_i[\sigma] = 1$ if and only if i plays every action on the path from the root to sequence $\sigma \in \Sigma_i$. We will use the shorthand $\boldsymbol{x}_i[z] = \boldsymbol{x}_i[\sigma_i(z)]$. A *mixed strategy* is a distribution over pure strategies, and the sequence form of a mixed strategy is the corresponding convex combination $\boldsymbol{x}_i \in [0, 1]^{\Sigma_i}$. We will use X_i to denote the polytope of sequence-form mixed strategies of player i.

A profile of mixed strategies $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \in X := X_1 \times \dots \times X_n$, induces a distribution over terminal nodes. We will use $z \sim \boldsymbol{x}$ to denote sampling from such a distribution. The expected utility of agent i under such a distribution is given by $u_i(\boldsymbol{x}) := \mathbb{E}_{z\sim\boldsymbol{x}}\, u_i(z)$. Critically, the sequence form has the property that each agent's expected utility is a linear function of its own sequence-form mixed strategy. For a profile $\boldsymbol{x} \in X$ and set $N \subseteq [\![n]\!]$, we will use the notation $\hat{\boldsymbol{x}}_N \in \mathbb{R}^Z$ to denote the vector $\hat{\boldsymbol{x}}_N[z] = \prod_{j\in N} \boldsymbol{x}_j[z]$, and we will write $\hat{\boldsymbol{x}} := \hat{\boldsymbol{x}}_{[\![n]\!]}$. A Nash equilibrium is a strategy profile $\boldsymbol{x}$ such that, for any $i \in [\![n]\!]$ and any $\boldsymbol{x}'_i \in X_i$, $u_i(\boldsymbol{x}) \ge u_i(\boldsymbol{x}'_i, \boldsymbol{x}_{-i})$.

3 The Steering Problem

In this section, we introduce what we call the *steering* problem. Informally, the steering problem asks whether a mediator can always steer players to any given equilibrium of an extensive-form game.

Definition 3.1 (Steering Problem for Pure-Strategy Nash Equilibrium)**.** Let Γ be an extensive-form game with payoffs bounded in $[0, 1]$. Let $\boldsymbol{d}$ be an arbitrary pure-strategy Nash equilibrium of Γ. The mediator knows the game Γ, as well as a function $R(T) = o(T)$, which may be game-dependent, that bounds the regret of all players. At each round $t \in [\![T]\!]$, the mediator picks *payment functions* for each player, $p_i^{(t)} : X_1 \times \dots \times X_n \to [0, P]$, where $p_i^{(t)}$ is linear in $\boldsymbol{x}_i$ and continuous in $\boldsymbol{x}_{-i}$, and P defines the largest allowable per-iteration payment. Then, players pick strategies $\boldsymbol{x}_i^{(t)} \in X_i$. Each player i then gets utility $v_i^{(t)}(\boldsymbol{x}_i) := u_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i}^{(t)}) + p_i^{(t)}(\boldsymbol{x}_i, \boldsymbol{x}_{-i}^{(t)})$. The mediator has two desiderata.

(S1) (Payments) The time-averaged realized payments to the players, defined as $\max_{i \in [\![n]\!]} \frac{1}{T} \sum_{t=1}^{T} p_i^{(t)}(\boldsymbol{x}^{(t)})$, converges to 0 as $T \to \infty$.

(S2) (Equilibrium) Players' actions are indistinguishable from the Nash equilibrium $\boldsymbol{d}$. That is, the *directness gap*, defined as $\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{z \sim \boldsymbol{x}^{(t)}} (1 - \hat{\boldsymbol{d}}[z])$, converges to 0 as $T \to \infty$.

The assumption imposed on the payment functions in Definition 3.1 ensures the existence of Nash equilibria in the payment-augmented game (*e.g.*, Glicksberg, 1952). Throughout this paper, we will refer to players as *direct* if they are playing actions prescribed by the target equilibrium strategy $\boldsymbol{d}$. Critically, (S2) does not require that the strategies themselves converge to the direct strategies, *i.e.*, $\boldsymbol{x}_i^{(t)} \to \boldsymbol{d}_i$, in iterates or in averages. They may differ on nodes off the equilibrium path. Instead, the requirement defined by (S2) is equivalent to *the reach probability of every node not reached in the equilibrium* $\mathbf{d}$ *converging to 0*, so that, *on path*, the players play the equilibrium. Similarly, (S1) refers to the *realized* payments $p_i^{(t)}(\boldsymbol{x}^{(t)})$, not the *maximum offered payment* $\max_{\boldsymbol{x} \in X} p_i^{(t)}(\boldsymbol{x})$.

For now, we will assume that a pure Nash equilibrium has been computed, and therefore our only task is to steer the agents toward it. In Section 6 we show how our steering algorithms can be directly applied to other equilibrium concepts such as *mixed* or *correlated* equilibria, and *communication* equilibria, and to the case where the equilibrium has not been precomputed.

The mediator does not know anything about how the players pick their strategies, except that they will have regret bounded by a function that vanishes in the limit and is known to the mediator. This condition is a commonly adopted behavioral assumption (Nekipelov et al., 2015; Kolumbus & Nisan, 2022b; Camara et al., 2020). The regret of Player $i \in [\![n]\!]$ in this context is defined as

$$\mathrm{Reg}_{X_i}^{T} := \frac{1}{P+1} \left[\max_{\boldsymbol{x}_i^* \in X_i} \sum_{t=1}^{T} v_i^{(t)}(\boldsymbol{x}_i^*) - \sum_{t=1}^{T} v_i^{(t)}(\boldsymbol{x}_i^{(t)}) \right].$$

That is, regret takes into account the payment functions offered to that player.[1] The assumption of bounded regret is realistic even in extensive-form games, as various regret minimizing algorithms exist. Two notable examples are the *counterfactual regret minimization* (CFR) framework (Zinkevich et al., 2007), which yields *full-feedback* regret minimizers, and IXOMD (Kozuno et al., 2021), which yields *bandit-feedback* regret minimizers.

How large payments are needed to achieve (S1) and (S2)? If the mediator could provide totally unconstrained payments, it could enforce any arbitrary outcome. On the other hand if the total payments are restricted to be bounded, the steering problem is information-theoretically impossible:

Proposition 3.2. *There exists a game and some function $R(T) = O(\sqrt{T})$ such that, for all $B \geq 0$, the steering problem is impossible if we add the constraint $\sum_{t=1}^{\infty} \sum_{i=1}^{n} p_i^{(t)}(\boldsymbol{x}^{(t)}) \leq B$.*

(Proofs are in Appendix E unless otherwise stated.) Hence, a weaker requirement on the size of the payments is needed. Between these extremes, one may allow the *total* payment to be unbounded, but insist that the *average* payment per round must vanish in the limit.

4 Steering in Normal-Form Games

We start with the example of *normal-form games*. A normal-form game, in our language, is simply an extensive-form game in which every player has one information set, and the set of histories correspond precisely to the set of pure profiles, *i.e.*, for every pure profile $\boldsymbol{x}$, we have $\hat{\boldsymbol{x}}[z] = 1$ for exactly one terminal node z. This setting is, much more simple than the general extensive-form setting which we will consider in the next section. In normal-form games, the strategy sets X_i are simplices, $X_i = \Delta(A_i)$, where A_i is the action set of player i at its only decision point. In this setting, we are able to turn to a special case of a result of Monderer & Tennenholtz (2004):

Theorem 4.1 (Costless implementation of pure Nash equilibria, special case of k-implementation, Monderer & Tennenholtz, 2004)**.** *Let $\mathbf{d}$ be a pure Nash equilibrium in a normal-form game. Then there exist functions $p_i^* : X_1 \times \cdots \times X_n \to [0, 1]$, with $p_i^*(\mathbf{d}) = 0$, such that in the game with utilities $v_i := u_i + p_i^*$, the profile $\mathbf{d}$ is weakly dominant: $v_i(\boldsymbol{d}_i, \boldsymbol{x}_{-i}) \geq v_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})$ for every profile $\boldsymbol{x}$.*

[1] The division by $1/(P+1)$ is for normalization, since $v_i^{(t)}$s has range $[0, P+1]$.

Indeed, it is easy to check that the payment function

$$p_i^*(\boldsymbol{x}) := (\boldsymbol{d}_i^\top \boldsymbol{x}_i)\Big(1 - \prod_{j \neq i} \boldsymbol{d}_j^\top \boldsymbol{x}_j\Big),$$

which on pure profiles $\boldsymbol{x}$ returns 1 if and only if $\boldsymbol{x}_i = \boldsymbol{d}_i$ and $\boldsymbol{x}_j \neq \boldsymbol{d}_j$ for some $j \neq i$, satisfies these properties. Such a payment function is *almost* enough for steering: the only problem is that $\boldsymbol{d}$ is only *weakly* dominant, so no-regret players *may* play other strategies than $\boldsymbol{d}$. This is easily fixed by adding a small reward $\alpha \ll 1$ for playing $\boldsymbol{d}_i$. That is, we set

$$p_i(\boldsymbol{x}) := \alpha \boldsymbol{d}_i^\top \boldsymbol{x}_i + p_i^*(\boldsymbol{x}) = (\boldsymbol{d}_i^\top \boldsymbol{x}_i)\Big(\alpha + 1 - \prod_{j \neq i} \boldsymbol{d}_j^\top \boldsymbol{x}_j\Big). \tag{1}$$

On a high level, the structure of the payment function guarantees that the average strategy of any no-regret learner $i \in [\![n]\!]$ should be approaching the direct strategy $\boldsymbol{d}_i$ by making $\boldsymbol{d}_i$ the strictly dominant strategy of player i. At the same time, it is possible to ensure that the average payment will also be vanishing by appropriately selecting parameter α. With appropriate choice of α, this is enough to solve the steering problem for normal-form games:

Theorem 4.2 (Normal-form steering)**.** *Let $p_i(\boldsymbol{x})$ be defined as in (1), set $\alpha = \sqrt{\varepsilon}$, where $\varepsilon := 4nR(T)/T$, and let T be large enough that $\alpha \leq 1$. Then players will be steered toward equilibrium, with both payments and directness gap bounded by $2\sqrt{\varepsilon}$.*

5 Steering in Extensive-Form Games

The extensive-form setting is significantly more involved than the normal-form setting, and it will be the focus for the remainder of our paper, for two reasons. First, in extensive form, the strategy spaces of the players are no longer simplices. Therefore, if we wanted to write a payment function p_i with the property that $p_i(\boldsymbol{x}) = \alpha \mathbb{1}\{\boldsymbol{x} = \boldsymbol{d}\} + \mathbb{1}\{\boldsymbol{x}_i = \boldsymbol{d}_i; \exists j\, \boldsymbol{x}_j \neq \boldsymbol{d}_j\}$ for pure $\boldsymbol{x}$ (which is what was needed by Theorem 4.2), such a function would not be linear (or even convex) in player i's strategy $\boldsymbol{x}_i \in X_i$ (which is a sequence-form strategy, not a distribution over pure strategies). As such, even the meaning of extensive-form regret minimization becomes suspect in this setting. Second, in extensive form, a desirable property would be that the mediator give payments conditioned only on what actually happens in gameplay, *not* on the players' full strategies—in particular, if a particular information set is not reached during play, the mediator should not know what action the player *would have* selected at that information set. We will call this the *bandit* setting, and distinguish it from the *full-feedback* setting, where the mediator observes the players' full strategies.[2] This distinction is meaningless in the normal-form setting: since terminal nodes in normal form correspond to (pure) profiles, observing gameplay is equivalent to observing strategies. (We will discuss this point in more detail when we introduce the bandit setting in Section 5.2.)

We now present two different algorithms for the steering problem, one in the full-feedback setting, and one in the bandit setting.

5.1 Steering with Full Feedback

In this section, we introduce a steering algorithm for extensive-form games under full feedback.

Algorithm 5.1 (FullFeedbackSteer)**.** At every round, set the payment function $p_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})$ as

$$\underbrace{\alpha \boldsymbol{d}_i^\top \boldsymbol{x}_i}_{\text{directness bonus}} + \underbrace{[u_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i}) - u_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})]}_{\text{sandboxing payments}} - \underbrace{\min_{\boldsymbol{x}_i' \in X_i} [u_i(\boldsymbol{x}_i', \boldsymbol{d}_{-i}) - u_i(\boldsymbol{x}_i', \boldsymbol{x}_{-i})]}_{\text{payment to ensure nonnegativity}}, \tag{2}$$

where $\alpha \leq 1/|Z|$ is a hyperparameter that we will select appropriately.

By construction, p_i satisfies the conditions of the steering problem (Definition 3.1): it is linear in $\boldsymbol{x}_i$, continuous in $\boldsymbol{x}_{-i}$, nonnegative, and bounded by an absolute constant (namely, 3). The payment function defined above has three terms:

[2] To be clear, the settings are differentiated by what the *mediator* observes, not what the *players* observe. That is, it is valid to consider the full-feedback steering setting with players running bandit regret minimizers, or the bandit steering setting with players running full-feedback regret minimizing algorithms.

1. The first term is a *reward for directness*: a player gets a reward proportional to α if it plays $\boldsymbol{d}_i$.
2. The second term *compensates the player* for the indirectness of other players. That is, the second term ensures that players' rewards are *as if* the other players had acted directly.
3. The final term simply ensures that the overall expression is nonnegative.

We claim that this protocol solves the basic version of the steering problem. Formally:

Theorem 5.2. *Set $\alpha = \sqrt{\varepsilon}$, where $\varepsilon := 4nR(T)/T$, and let T be large enough that $\alpha \le 1/|Z|$. Then,* FULLFEEDBACKSTEER *results in average realized payments and directness gap at most $3|Z|\sqrt{\varepsilon}$.*

5.2 Steering in the Bandit Setting

In FULLFEEDBACKSTEER, payments depend on full strategies $\boldsymbol{x}$, not the realized game trajectories. In particular, the mediator in Theorem 5.2 observes what the players *would have played* even at infosets that other players avoid. To allow for an algorithm that works without knowledge of full strategies, $p_i^{(t)}$ must be structured so that it could be induced by a payment function that only gives payments for terminal nodes reached during play. To this end, we now formalize *bandit steering*.

Definition 5.3 (Bandit steering problem)**.** Let Γ be an extensive-form game in which rewards are bounded in $[0, 1]$ for all players. Let $\boldsymbol{d}$ be an arbitrary pure-strategy Nash equilibrium of Γ. The mediator knows Γ and a regret bound $R(T) = o(T)$. At each $t \in [\![T]\!]$, the mediator selects a payment function $q_i^{(t)} : Z \to [0, P]$. The players select strategies $\boldsymbol{x}_i^{(t)}$. A terminal node $z^{(t)} \sim \boldsymbol{x}^{(t)}$ is sampled, and all agents observe the terminal node that was reached, $z^{(t)}$. The players get payments $q_i^{(t)}(z^{(t)})$, so that their expected payment is $p_i^{(t)}(\boldsymbol{x}) := \mathbb{E}_{z\sim\boldsymbol{x}}\, q_i^{(t)}(z)$. The desiderata are as in Definition 3.1.

The bandit steering problem is more difficult than the non-bandit steering problem in two ways. First, as discussed above, the mediator does not observe the strategies $\boldsymbol{x}$, only a terminal node $z^{(t)} \sim \boldsymbol{x}$. Second, the form of the payment function $q_i^{(t)} : Z \to [0, P]$ is restricted: this is already sufficient to rule out FULLFEEDBACKSTEER. Indeed, p_i as defined in (2) cannot be written in the form $\mathbb{E}_{z\sim\boldsymbol{x}}\, q_i(z)$: $p_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})$ is nonlinear in $\boldsymbol{x}_{-i}$ due to the nonnegativity-ensuring payments, whereas every function of the form $\mathbb{E}_{z\sim\boldsymbol{x}}\, q_i(z)$ will be linear in each player's strategy.

We remark that, despite the above algorithm containing a sampling step, the payment function is defined *deterministically*: the payment is defined as the *expected value* $p_i^{(t)}(\boldsymbol{x}) := \mathbb{E}_{z\sim\boldsymbol{x}}\, q_i^{(t)}(z)$. Thus, the theorem statements in this section will also be deterministic.

In the normal-form setting, the payments p_i defined by (1) already satisfy the condition of bandit steering. In particular, let z be the terminal node we have $p_i(\boldsymbol{x}) = \mathbb{E}_{z\sim\boldsymbol{x}}\left[\alpha\mathbb{1}\{z = z^*\} + \mathbb{1}\{\boldsymbol{x}_i = \boldsymbol{d}_i; \exists j\, \boldsymbol{x}_j \neq \boldsymbol{d}_j\}\right]$. Therefore, in the normal-form setting, Theorem 4.2 applies to both full-feedback steering and bandit steering, and we have no need to distinguish between the two. However, in extensive form, as discussed above, the two settings are quite different.

5.2.1 Lower Bound on Required Payments

Unlike in the full-feedback or normal-form settings, in the bandit setting, steering is impossible in the general case in the sense that per-iteration payments bounded by any constant do not suffice.

Theorem 5.4. *For every $P > 0$, there exists an extensive-form game Γ with $O(P)$ players, $O(P^2)$ nodes, and rewards bounded in $[0, 1]$ such that, with payments $q_i^{(t)} : Z \to [0, P]$, it is impossible to steer players to the welfare-maximizing Nash equilibrium, even when $R(T) = 0$.*

For intuition, consider the extensive-form game in Figure 2, which can be seen as a three-player version of Stag Hunt. Players who play Hare (H) get a value of $1/2$ (up to constants); in addition, if all three players play Stag (S), they all get expected value 1. The welfare-maximizing equilibrium is "everyone plays Stag", but "everyone plays Hare" is also an equilibrium. In addition, if all players are playing Hare, the only way for the mediator to convince a player to play Stag without accidentally also paying players in the Stag equilibrium is to pay players at one of the three boxed nodes. But those three nodes are only reached with probability $1/n$ as often as the three nodes on the left, so the mediator would have to give a bonus of more than $n/2$. The full proof essentially works by deriving an algorithm that the players could use to exploit this dilemma to achieve either large payments or bad convergence rate, generalizing the example to $n > 3$, and taking $n = \Theta(P)$.

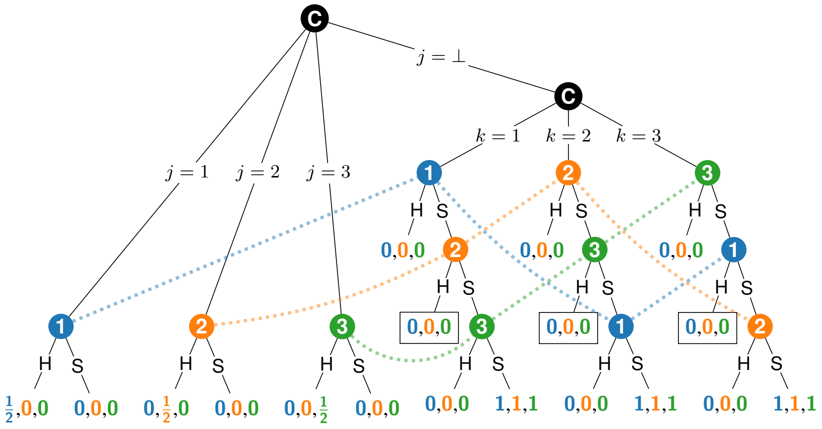


Figure 2: The counterexample for Theorem 5.4, for $n = 3$. Chance always plays uniformly at random. Infosets are linked by dotted lines (all nodes belonging to the same player are in the same infoset).

5.2.2 Bandit Steering with Large Off-Path Payments

To circumvent the lower bound in Theorem 5.4, in this subsection, we allow the payment bound $P \geq 1$ to depend on both the time limit T and the game. Consider the following algorithm.

Algorithm 5.5 (BanditSteer)**.** Let α, P be hyperparameters. Then, for all rounds $t = 1, \dots, T$, sample $z \sim \boldsymbol{x}^{(t)}$ and pay players as follows. If all players have been direct (*i.e.*, if $\hat{\boldsymbol{d}}[z] = 1$), pay all players α. If at least one player has not been direct, pay P to all players who have been direct. That is, set $q_i^{(t)}(z^{(t)}) = \alpha \hat{\boldsymbol{d}}[z] + P \boldsymbol{d}_i[z](1 - \hat{\boldsymbol{d}}[z])$.

Theorem 5.6. *Set the hyperparameters $\alpha = 4|Z|^{1/2}\varepsilon^{1/4}$ and $P = 2|Z|^{1/2}\varepsilon^{-1/4}$, where $\varepsilon := R(T)/T$, and let T be large enough that $\alpha \leq 1$. Then, running* BanditSteer *for T rounds results in average realized payments bounded by $8|Z|^{1/2}\varepsilon^{1/4}$, and directness gap by $2\varepsilon^{1/2}$.*

The proof of this result is more involved than those for previous results. One may hope that—as in FullFeedbackSteer—the desired equilibrium can be made dominant by adding payments. But this is impossible: in the smaller "stag hunt" game in Figure 1, for Player 2, Stag cannot be a weakly-dominant strategy unless a payment is given at the boxed node, which would be problematic because such payments would also appear in equilibrium, in violation of (S1). In fact, a sort of "chicken-and-egg" problem arises: (S2) requires that all players converge to equilibrium. But for this to happen, other players' strategies must first converge to equilibrium so that i's incentives are as they would be in equilibrium. The main challenge in the proof of Theorem 5.6 is therefore to carefully set the hyperparameters to achieve convergence despite these apparent problems.

6 Other Equilibrium Notions and Online Steering

So far, Theorems 5.2 and 5.6 refer only to *pure-strategy* Nash equilibria of a game. We now show how to apply these algorithms to other equilibrium notions such as mixed-strategy or correlated equilibrium. The key insight is that many types of equilibrium can be viewed as pure-strategy equilibria in an augmented game. For example, an extensive-form correlated equilibrium of a game Γ can be viewed as a pure-strategy equilibrium of an augmented game Γ' in which the mediator samples actions ("recommendations") and the acting player observes those recommendations. Then, in Γ', the goal is to guide players toward the pure strategy profile of following recommendations.

We now formalize these ideas. For this section, let Γ refer to a *mediator-augmented* game (Zhang & Sandholm, 2022), which has $n + 1$ players $i \in [\![n]\!] \cup \{0\}$, where player 0 is the mediator. We will assume the *revelation principle*, which allows us to fix a target pure strategy profile $\boldsymbol{d}$ that we want to make the equilibrium profile for the non-mediator players. We will write $\Gamma^{\boldsymbol{\mu}}$ to refer to the n-player game in which the mediator is fixed to playing the strategy $\boldsymbol{\mu}$.

Definition 6.1. An *equilibrium in the mediator-augmented game* Γ is a strategy $\boldsymbol{\mu} \in X_0$ for the mediator such that $\boldsymbol{d}$ is a Nash equilibrium of $\Gamma^{\boldsymbol{\mu}}$. An equilibrium $\boldsymbol{\mu}$ is *optimal* if, among all equilibria, it maximizes the mediator's objective $u_0(\boldsymbol{\mu}, \boldsymbol{d})$.

By varying the construction of the augmented game Γ, the family of solution concepts for extensive-form games captured by this framework includes, but is not limited to, normal-form coarse correlated equilibrium (Aumann, 1974; Moulin & Vial, 1978); extensive-form correlated equilibrium (EFCE)[3] (von Stengel & Forges, 2008); communication equilibrium (Forges, 1986); mechanism design; and information design/Bayesian persuasion (Kamenica & Gentzkow, 2011).

Unlike the offline setting (where the target equilibrium is given to us), in the online setting we can choose the target equilibrium. In particular, we would like to steer players toward an *optimal* equilibrium $\boldsymbol{\mu}$, without knowing that equilibrium beforehand. To that end, we add a new criterion:

(S3) (Optimality) The mediator's reward should converge to the reward of the optimal equilibrium. That is, the *optimality gap* $u_0^* - \frac{1}{T}\sum_{t=1}^{T} u_0(\boldsymbol{\mu}^{(t)}, \boldsymbol{x}^{(t)})$, where u_0^* is the mediator utility in an optimal equilibrium, converges to 0 as $T \to \infty$.

In Appendix D, we discuss why it is in some sense necessary to allow the mediator to give recommendations, not just payments, if the target equilibrium is not pure.

Since equilibria in mediator-augmented games are just strategies $\boldsymbol{\mu}$ under which $\boldsymbol{d}$ is a Nash equilibrium, we may use the following algorithm to steer players toward an optimal equilibrium of Γ:

Algorithm 6.2 (COMPUTETHENSTEER)**.** Compute an optimal equilibrium $\boldsymbol{\mu}$. With $\boldsymbol{\mu}$ held fixed, run any steering algorithm in $\Gamma^{\boldsymbol{\mu}}$.

As observed earlier, the main weakness of COMPUTETHENSTEER is that it must compute an equilibrium offline. To sidestep this, in this section we will introduce algorithms that compute the equilibrium in an *online* manner, while steering players toward it. Our algorithms will make use of a Lagrangian dual formulation analyzed by Zhang et al. (2023).

Proposition 6.3 (Zhang et al. (2023))**.** *There exists a (game-dependent) constant $\lambda^* \geq 0$ such that, for every $\lambda \geq \lambda^*$, the solutions $\boldsymbol{\mu}$ to*

$$\max_{\boldsymbol{\mu} \in X_0} \min_{\boldsymbol{x} \in X} u_0(\boldsymbol{\mu}, \boldsymbol{d}) - \lambda \sum_{i=1}^{n} [u_i(\boldsymbol{\mu}, \boldsymbol{x}_i, \boldsymbol{d}_{-i}) - u_i(\boldsymbol{\mu}, \boldsymbol{d}_i, \boldsymbol{d}_{-i})], \tag{3}$$

are exactly the optimal equilibria of the mediator-augmented game.

Algorithm 6.4 (ONLINESTEER)**.** The mediator runs a regret minimization algorithm $\mathcal{R}_0$ over its own strategy space X_0, which we assume has regret at most $R_0(T)$ after T rounds. On each round, the mediator does the following:

- Get a strategy $\boldsymbol{\mu}^{(t)}$ from $\mathcal{R}_0$. Play $\boldsymbol{\mu}^{(t)}$, and set $p_i^{(t)}$ as defined in (2) in $\Gamma^{\boldsymbol{\mu}^{(t)}}$.
- Pass utility $\boldsymbol{\mu} \mapsto \frac{1}{\lambda} u_0(\boldsymbol{\mu}, \boldsymbol{d}) - \sum_{i=1}^{n} \left[u_i(\boldsymbol{\mu}, \boldsymbol{x}_i^{(t)}, \boldsymbol{d}_{-i}) - u_i(\boldsymbol{\mu}, \boldsymbol{d}_i, \boldsymbol{d}_{-i}) \right]$ to $\mathcal{R}_0$, where $\lambda \geq 1$ is a hyperparameter.

Theorem 6.5. *Set the hyperparameters $\alpha = \varepsilon^{2/3}|Z|^{-1/3}$ and $\lambda = |Z|^{2/3}\varepsilon^{-1/3}$, where $\varepsilon := (R_0(T) + 4nR(T))/T$ is the average regret bound summed across players, and let T be large enough that $\alpha \leq 1/|Z|$. Then running* ONLINESTEER *results in average realized payments, directness gap, and optimality gap all bounded by $7\lambda^*|Z|^{4/3}\varepsilon^{1/3}$.*

The argument now works with the zero-sum formulation (3), and leverages the fact that the agents' average strategies are approaching the set of Nash equilibria since they have vanishing regrets. Thus, each player's average strategy should be approaching the direct strategy, which in turn implies that the average utility of the mediator is converging to the optimal value, analogously to Theorem 5.2.

ONLINESTEER has a further guarantee that FULLFEEDBACKSTEER does not, owing to the fact that it learns an equilibrium online: it works even when the players' sets of deviations, X_i, is not known upfront. In particular, the following generalization of Theorem 6.5 follows from an identical proof.

[3] This requires the mediator to have imperfect recall.

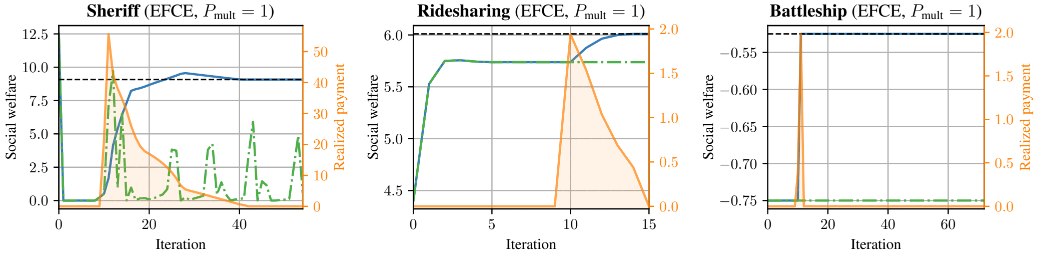


Figure 3: Sample experimental results. The blue line in each figure is the social welfare (left y-axis) of the players *with* steering enabled. The green dashed line is the social welfare *without* steering. The yellow line gives the payment (right y-axis) paid to each player. The flat black line denotes the welfare of the optimal equilibrium. The panels show the game, the equilibrium concept (in this figure, always EFCE). In all cases, the first ten iterations are a "burn-in" period during which no payments are issued; steering only begins after that.

Corollary 6.6. *Suppose that each player i, unbeknownst to the mediator, is choosing from a subset $Y_i \subseteq X_i$ of strategies that includes the direct strategy $\mathbf{d}_i$. Then, running Theorem 6.5 with the same hyperparameters yields the same convergence guarantees, except that the mediator's utility converges to its optimal utility against the* true *deviators, that is, a solution to (3) with each X_i replaced by Y_i.*

At this point, it is very reasonable to ask whether it is possible to perform *online* steering with *bandit* feedback. In *normal-form* games, as with offline setting, there is minimal difference between the bandit and the full-feedback setting. This intuition carries over to the bandit setting: ONLINESTEER can be adapted into an online bandit steering algorithm for normal-form games, with essentially the same convergence guarantee. We defer the formal statement of the algorithm and proof to Appendix F.

The algorithm, however, fails to extend to the *extensive-form* online bandit setting, for the same reasons that the *offline* full-feedback algorithm fails to extend to the online setting.

7 Experimental Results

We ran experiments with our BANDITSTEER algorithm (Algorithm 5.5) on various notions of equilibrium in extensive-form games, using the COMPUTETHENSTEER framework suggested by Algorithm 6.2. Since the hyperparameter settings suggested by Algorithm 5.5 are very extreme, in practice we fix a constant P and set α dynamically based on the currently-observed gap to directness. We used CFR+ (Tammelin, 2014) as the regret minimizer for each player, and precomputed a welfare-optimal equilibrium with the LP algorithm of Zhang & Sandholm (2022). In most instances tested, a small constant P (say, $P \leq 8$) is enough to steer CFR+ regret minimizers to the exact equilibrium in a finite number of iterations. Two plots exhibiting this behavior are shown in Figure 3. More experiments, as well as descriptions of the game instances tested, can be found in Appendix G.

8 Conclusions and Future Research

We established that it is possible to steer no-regret learners to optimal equilibria using vanishing rewards, even under bandit feedback. There are many interesting avenues for future research. First, is there a natural *bandit online* algorithm that combines the desirable properties of both ONLINESTEER and BANDITSTEER? Also, it is important to understand the best rates attainable for the different settings of the steering problem. Furthermore, is there a steering algorithm for which the mediator needs to know even less information about the game upfront? For example, could a mediator without knowledge of the players' utilities still steer toward optimal equilibria? Finally, our main behavioral assumption throughout this paper is that players incur vanishing average regret. Yet, stronger guarantees are possible when specific no-regret learning dynamics are in place; *e.g.*, see (Vlatakis-Gkaragkounis et al., 2020; Giannou et al., 2021a;b) for recent characterizations in the presence of *strict* equilibria. Concretely, it would be interesting to understand the class of learning dynamics under which the steering problem can be solved with a finite cumulative budget.

REFERENCES

Lucas Agussurja and Hoong Chuin Lau. The price of stability in selfish scheduling games. *Web Intell. Agent Syst.*, 7(4):321–332, 2009.

Elliot Anshelevich, Anirban Dasgupta, Jon M. Kleinberg, Éva Tardos, Tom Wexler, and Tim Roughgarden. The price of stability for network design with fair cost allocation. *SIAM Journal on Computing*, 38(4):1602–1623, 2008.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32:48–77, 2002.

Robert Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1:67–96, 1974.

Maria-Florina Balcan. Leading dynamics to good behavior. *SIGecom Exch.*, 10(2):19–22, 2011.

Maria-Florina Balcan, Avrim Blum, and Yishay Mansour. Improved equilibria via public service advertising. In *Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 728–737, 2009.

Maria-Florina Balcan, Avrim Blum, and Yishay Mansour. Circumventing the price of anarchy: Leading dynamics to good behavior. *SIAM Journal on Computing*, 42(1):230–264, 2013.

Maria-Florina Balcan, Sara Krehbiel, Georgios Piliouras, and Jinwoo Shin. Near-optimality in covering games by exposing global information. *ACM Trans. Economics and Comput.*, 2(4): 13:1–13:22, 2014.

Mark Braverman, Jieming Mao, Jon Schneider, and Matt Weinberg. Selling to a no-regret buyer. In *ACM Conference on Economics and Computation (EC)*, pp. 523–538, 2018.

William Brown, Jon Schneider, and Kiran Vodrahalli. Is learning in games good for the learners? *CoRR*, abs/2305.19496, 2023.

Linda Cai, S. Matthew Weinberg, Evan Wildenhain, and Shirley Zhang. Selling to multiple no-regret buyers. *CoRR*, abs/2307.04175, 2023.

Modibo K Camara, Jason D Hartline, and Aleck Johnsen. Mechanisms for a no-regret agent: Beyond the common prior. In *Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 259–270. IEEE, 2020.

In-Koo Cho and Jonathan Libgober. Machine learning for strategic inference. *arXiv preprint arXiv:2101.09613*, 2021.

Maurizio D'Andrea. Playing against no-regret players. *Operations Research Letters*, 2023.

Yuan Deng, Pingzhong Tang, and Shuran Zheng. Complexity and algorithms of k-implementation. In *Autonomous Agents and Multi-Agent Systems*, pp. 5–13. ACM, 2016.

Yuan Deng, Jon Schneider, and Balasubramanian Sivan. Strategizing against no-regret learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1577–1585, 2019.

Paul Duetting, Tomer Ezra, Michal Feldman, and Thomas Kesselheim. Multi-agent contracts. *CoRR*, abs/2211.05434, 2022.

Paul Dütting, Tomer Ezra, Michal Feldman, and Thomas Kesselheim. Combinatorial contracts. In *Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 815–826. IEEE, 2021a.

Paul Dütting, Tim Roughgarden, and Inbal Talgam-Cohen. The complexity of contracts. *SIAM Journal on Computing*, 50(1):211–254, 2021b.

Gabriele Farina, Chun Kai Ling, Fei Fang, and Tuomas Sandholm. Correlation in extensive-form games: Saddle-point formulation and benchmarks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Francoise Forges. An approach to communication equilibria. *Econometrica: Journal of the Econometric Society*, pp. 1375–1385, 1986.

Rupert Freeman, David M. Pennock, Chara Podimata, and Jennifer Wortman Vaughan. No-regret and incentive-compatible online learning. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3270–3279. PMLR, 2020.

Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 22655–22666, 2021a.

Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2147–2148. PMLR, 2021b.

I. L. Glicksberg. A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. *American Mathematical Society*, 3(1):170–174, 1952.

Shengyuan Hu, Dung Daniel Ngo, Shuran Zheng, Virginia Smith, and Zhiwei Steven Wu. Federated learning as a network effects game, 2023.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011.

David Kempe, Sixie Yu, and Yevgeniy Vorobeychik. Inducing equilibria in networked public goods games through network structure modification. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20*, pp. 611–619. International Foundation for Autonomous Agents and Multiagent Systems, 2020.

Robert D. Kleinberg, Katrina Ligett, Georgios Piliouras, and Éva Tardos. Beyond the nash equilibrium barrier. In *Innovations in Computer Science - ICS 2011*, pp. 125–140. Tsinghua University Press, 2011.

Yoav Kolumbus and Noam Nisan. How and why to manipulate your own agent: On the incentives of users of learning agents. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022a.

Yoav Kolumbus and Noam Nisan. Auctions between regret-minimizing agents. In Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini (eds.), *WWW '22: The ACM Web Conference 2022*, pp. 100–111. ACM, 2022b.

Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. In *Symposium on Theoretical Aspects in Computer Science*, 1999.

Tadashi Kozuno, Pierre Ménard, Remi Munos, and Michal Valko. Learning in two-player zero-sum partially observable markov games with perfect recall. *Advances in Neural Information Processing Systems*, 34:11987–11998, 2021.

H. W. Kuhn. A simplified two-person poker. In H. W. Kuhn and A. W. Tucker (eds.), *Contributions to the Theory of Games*, volume 1 of *Annals of Mathematics Studies, 24*, pp. 97–103. Princeton University Press, Princeton, New Jersey, 1950.

Jiayang Li, Jing Yu, Yu Marco Nie, and Zhaoran Wang. End-to-end learning and intervention in games. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Kai Li, Wenhan Huang, Chenchen Li, and Xiaotie Deng. Exploiting a no-regret opponent in repeated zero-sum games. *Journal of Shanghai Jiaotong University (Science)*, 2023.

Boyi Liu, Jiayang Li, Zhuoran Yang, Hoi-To Wai, Mingyi Hong, Yu Nie, and Zhaoran Wang. Inducing equilibria via incentives: Simultaneous design-and-play ensures global convergence. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Yishay Mansour, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan. Strategizing against learners in bayesian games. In *Conference on Learning Theory (COLT)*, volume 178 of *Proceedings of Machine Learning Research*, pp. 5221–5252. PMLR, 2022.

David Mguni, Joel Jennings, Emilio Sison, Sergio Valcarcel Macua, Sofia Ceppi, and Enrique Munoz de Cote. Coordinating the crowd: Inducing desirable equilibria in non-cooperative systems. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pp. 386–394. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

Dov Monderer and Moshe Tennenholtz. K-implementation. *J. Artif. Intell. Res.*, 21:37–62, 2004.

H. Moulin and J.-P. Vial. 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, 1978.

Roger B Myerson. Multistage games with communication. *Econometrica: Journal of the Econometric Society*, pp. 323–358, 1986.

Denis Nekipelov, Vasilis Syrgkanis, and Éva Tardos. Econometrics for learning agents. In *ACM Conference on Economics and Computation (EC)*, pp. 1–18, 2015.

Ioannis Panageas and Georgios Piliouras. Average case performance of replicator dynamics in potential games via computing regions of attraction. In *ACM Conference on Economics and Computation (EC)*, pp. 703–720. ACM, 2016.

Tim Roughgarden. *Selfish routing and the price of anarchy*. MIT Press, 2005.

Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM*, 62(5):32:1–32:42, 2015.

Tim Roughgarden and Okke Schrijvers. Online prediction with selfish experts. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1300–1310, 2017.

Andreas S. Schulz and Nicolás E. Stier Moses. On the performance of user equilibria in traffic networks. In *Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 86–87, 2003.

Oskari Tammelin. Solving large imperfect information games using CFR+. *arXiv preprint arXiv:1407.5042*, 2014.

Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Thanasis Lianeas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed nash equilibria: They do not mix. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

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

Brian Hu Zhang and Tuomas Sandholm. Polynomial-time optimal equilibria with a mediator in extensive-form games. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Brian Hu Zhang, Gabriele Farina, Andrea Celli, and Tuomas Sandholm. 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)*, pp. 1119–1120. ACM, 2022.

Brian Hu Zhang, Gabriele Farina, Ioannis Anagnostides, Federico Cacciamani, Stephen McAleer, Andreas Haupt, Andrea Celli, Nicola Gatti, Vincent Conitzer, and Tuomas Sandholm. Computing optimal equilibria and mechanisms via learning in zero-sum extensive-form games. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Martin Zinkevich, Michael Bowling, Michael Johanson, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2007.

A Related Work

Our setting and algorithms are closely related to the problem of *k-implementation* (Monderer & Tennenholtz, 2004) in normal-form games (see also (Deng et al., 2016) for pertinent complexity considerations). In k-implementation, the goal is to make a certain strategy profile a *(weakly) dominant strategy* for all players using nonnegative payments. Monderer & Tennenholtz (2004) observe that only Nash equilibria can be implemented using zero realized payments. Our FULLFEEDBACKSTEER algorithm operates in a similar setting: by precomputing an equilibrium and giving payments in such a way that players are *sandboxed*, each player's dominant strategy is to be direct, so the players converge. Indeed, for *normal-form* games, the fact that all pure Nash equilibria are (in the language of k-implementation) 0-implementable implies that steering is possible for normal-form games, in both the full-feedback and online settings. Our FULLFEEDBACKSTEER and BANDITSTEER algorithms could then be interpreted as saying that arbitrary Nash equilibria of extensive-form games can be implemented in unique normal-form coarse correlated equilibria (and therefore the unique convergence point of no-regret learning dynamics). However, our results differ from k-implementation in a few crucial ways: (1) Our rationality assumption differs: Our algorithms seek to steer *no-regret* learners, instead of players that play weakly-dominant strategies. Indeed, in the bandit setting, we argued earlier that, in extensive form, it is sometimes impossible to make the desirable equilibrium weakly dominant, and this leads to a more intricate proof for Theorem 5.6. (2) We consider a wider class of games: Our algorithms work in arbitrary extensive-form settings, not just normal form. As we discussed above, this causes unique problems in the bandit setting. Even in the full-information setting, working in extensive form means that we need to be careful in designing the payment scheme so that the maximum possible payment P is constant. For instance, Theorem 5.4 shows that no absolute constant payment can suffice. (3) We restrict the information available to the mediator: Algorithm ONLINESTEER *learns the equilibrium while steering agents toward it*.

Moreover, ample of prior research has endeavored to steer strategic agents toward "good" equilibria (Mguni et al., 2019; Li et al., 2020; Kempe et al., 2020; Liu et al., 2022). Indeed, the presence of a centralized party that can help "nudge" behavior to a better state has served as a central motivation for the literature on the *price of stability* (Anshelevich et al., 2008; Schulz & Moses, 2003; Agussurja & Lau, 2009; Panageas & Piliouras, 2016), thereby allowing to circumvent impossibility results in terms of the worst Nash equilibria (Koutsoupias & Papadimitriou, 1999; Roughgarden, 2005). For example, as articulated by Balcan et al. (2009): "In cases where there are both high and low cost Nash equilibria, a central authority could hope to "move" behavior from a high-cost equilibrium to a low-cost one by running a public service advertising campaign promoting the better behavior." Nevertheless, Balcan et al. (2009) also stress that it is unrealistic to assume that all agents blindly follow the prescribed protocol, unless it is within their interest to do so; this is indeed a key motivation for our considerations. Balcan (2011); Balcan et al. (2013; 2014) also endeavor to lead learning dynamics to a desired state for certain classes of games, although there are key differences between those papers and our setting. In particular, focusing on the work of Balcan et al. (2014) for concreteness, our paper shows that steering is possible under the mild assumption that players have vanishing average regret, while Balcan et al. (2014) impose much stronger behavioral assumptions; namely, in the first phase of their protocol players who receive advise are assumed to obey, even though it may not be in their own interest, while the rest of the players are following best response dynamics. Further, while in the protocol of Balcan et al. (2014) advise is provided to a subset of the players, they only guarantee convergence to an *approximately* optimal state; by contrast, our focus here is on steering to optimal equilibria.

On a related direction, Kleinberg et al. (2011) identify a class of games where specific learning dynamics lead to much better social welfare compared to the Nash equilibrium. More broadly, Roughgarden's smoothness framework (Roughgarden, 2015) gives bounds on the (time-average) social welfare guarantees under no-regret learners, but imposes somewhat restrictive assumptions on the underlying class of games.

Our problem of steering no-regret learners to desirable outcomes is also somewhat connected to the problem of strategizing against no-regret learners, studied from different perspectives in several prior papers (Deng et al., 2019; Kolumbus & Nisan, 2022a; Freeman et al., 2020; Roughgarden & Schrijvers, 2017; D'Andrea, 2023; Cho & Libgober, 2021; Mansour et al., 2022; Brown et al., 2023; Li et al., 2023; Cai et al., 2023). We elaborate now on the connection in particular with the results of

Deng et al. (2019). The setting of that paper is quite different from the one that we study, in several important ways.

- Their setting is a *normal-form* setting with *one leader* and *one follower*, in which a critical assumption is that the follower has no weakly dominated pure strategies. We make none of these assumptions: our setting is an *extensive form* setting with *multiple followers* in general, and we make no assumptions about domination. These changes make the problem significantly harder: in fact, in their setting, payments are not even required to guide the learner to the "equilibrium" (best response); it suffices for the leader to be clever about picking its mixed strategy. In our setting, however, payments are most certainly required for steering, even in the one-player case.

- Our mediator is more restricted in terms of what it can do to affect actual gameplay: our mediator can only *talk to* and *pay* the players; it cannot directly affect the outcome of the game in the way that a Stackelberg leader can, since the players are free to ignore or disobey the mediator's actions. If the mediator wants something to happen, it must actually incentivize the players to do it.

Moreover, introducing nonnegative payments to incentivize specific outcomes bears resemblance to the setting of *contract design* (Duetting et al., 2022; Dütting et al., 2021b;a), and has been recently employed in *federated learning* as well to encourage participation (Hu et al., 2023). Finally, our study relates to the literature of mechanism design that adopts vanishing regret as a behavioral assumption (Camara et al., 2020; Braverman et al., 2018).

B Adapting the Framework of Zhang and Sandholm [2022]

In this section, we give some more detail on the framework of Zhang & Sandholm (2022) and how we adapt it in our paper.

Zhang & Sandholm (2022) base their framework on the notion of *communication equilibrium* (Forges, 1986; Myerson, 1986). In particular, they start with a *base game*, which we will call Γ_0, that has no mediator. The base game, together with a choice of notion of equilibrium, defines a *mediator-augmented game*, which is what we call Γ.

In this setup, if we wish our algorithms to run efficiently *in the size of the base game*, care must be taken in constructing the game Γ from Γ_0. In particular, if we try to build Γ by naively allowing the mediator to have one round of communication with the acting player at every information set before they act, the size of Γ blows up exponentially. To circumvent this, Zhang & Sandholm (2022) make several simplifications, all of which they show to be without loss of generality due to the revelation principle:

(M1) Players are not allowed to send messages that *prove* that they themselves have deviated from the direct strategy—for example, messages that are not information reports, or messages that are inconsistent with the past communication transcript between that player and the mediator. Players are instead allowed to send no message.

(M2) If a player sends no message, the mediator sends no action recommendation.

(M3) After one player deviates, all other players are assumed to play honestly.

With these simplifications, Γ has $O(|H||\Sigma|)$ nodes, where H and Σ are the sets of nodes and sequences in Γ_0, respectively. As such, solving the induced linear program takes time polynomial in $|H|$ and the description of the mediator's strategy space Ξ.

We now turn to our paper. When our goal is merely to *compute a good equilibrium*, it suffices to use this game Γ. However, when the goal is to *steer* players toward equilibria, as Section 3, Assumption (M3) becomes problematic: since we cannot control the players, we cannot assume they will play directly in the face of other deviations.

To repair this problem, we split the analysis into two different cases.

When the mediator has perfect recall, we use different assumption:

(M3') After the mediator proves that at least two players have deviated, the mediator announces this to all players and then ceases all communication with all players.

This creates a new mediator-augmented game Γ'. This game clearly has all the nodes of Γ and more, and no node in Γ' but not in Γ is ever touched by any strategy profile in which only a single player is not direct. Therefore, the change from (M3) to (M3') is equilibrium-preserving. Further, each history in Γ' can be identified by a tuple $(h_1, h_2, \sigma_1, \sigma_2)$, where h_1 is a history in Γ_0, and $(h_2, \sigma_1, \sigma_2)$ describes the mediator's transcripts with the players: the mediator sees the transcript that would be created if the true history were h_2 and at most two players σ_1, σ_2 have deviated. As such, Γ' has at most $O(|H|^2|\Sigma|^2)$ nodes, which is still polynomial in the size of the original game.

When the mediator does not have perfect recall—which is the case for the notions of *correlated* equilibria—the condition (M3') no longer makes sense, because an imperfect-recall agent can, by definition, forget things. As such, in these cases, we simply drop (M3) completely. In this case, a history in Γ' is identified by a tuple (h, σ) where $h \in H$, $\sigma \in \bigtimes_i \Sigma_i$. So, Γ' in this case has size $O(|H| \prod_i |\Sigma_i|) \le O(|H||\Sigma|^n)$. While this is exponential in the original game's size, we stress that some exponential blow-up is to some extent unavoidable already in the imperfect-recall setting: computing optimal NFCCEs, EFCCEs, and EFCEs are, after all, NP-hard.

In Section 6, we always implicitly operate on the game Γ' instead of Γ. This allows us to avoid issues of exponential blowup in the game size when it is avoidable.

C Details on Figure 1

In this section, we elaborate on Figure 1, and we provide some further pertinent illustrations. As shown in Theorem 5.4, this is a challenging instance for steering no-regret learners in the bandit setting. The results illustrated in Figure 1 correspond to each player employing multiplicative weights update (MWU) under full feedback with learning rate $\eta := 0.1$.

Furthermore, we also experiment with each player using a variant of EXP3 (Auer et al., 2002) with exploration parameter $\epsilon := 5\%$. We employ our steering algortihm in the bandit setting with different *potential* payments P and parameter $\alpha = 0$, leading to the results illustrated in Figure 4.

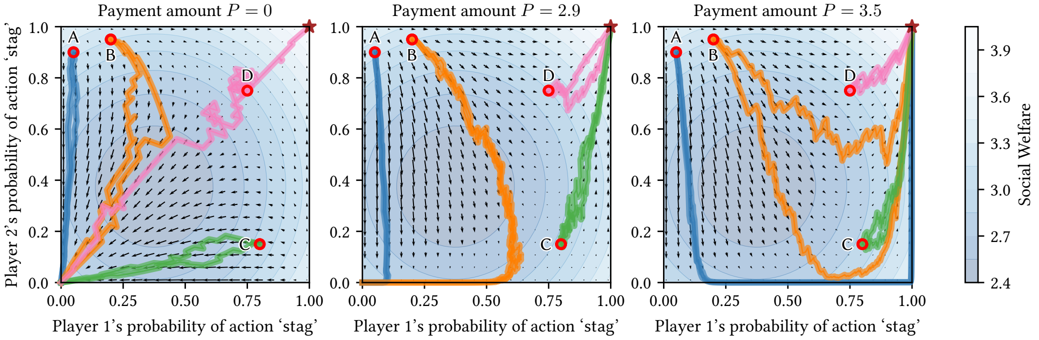


Figure 4: The trajectories of bandit algorithms under different random initializations and vanishing payments. Trajectories with the same color correspond to the same initialization but under different realizations of the players' sampled actions.

D Necessity of Recommendations for Mixed Equilibria

In this section, we discuss why recommendations are in some sense necessary for steering to equilibrium concepts other than pure Nash.[4]

For simplicity of exposition in the body and cleanliness of the proofs, our desiderata for steering are *defined* with pure equilibria in mind. In particular, Item (S2) states that, for all terminal nodes z for which $\hat{\boldsymbol{d}}[z] \neq 1$ (*i.e.*, for which at least one player i does not deterministically play to z), z will not be played by the players in the long run. Clearly, this definition does not make sense when the direct strategy is not pure. So, we must first define what it means to steer to a mixed equilibrium at all. We define:

(S2') Players' actions are indistinguishable from the Nash equilibrium $\boldsymbol{d}$, in the sense that $\hat{\boldsymbol{x}}^{(t)}[z] \to \hat{\boldsymbol{d}}[z]$ for all terminal nodes z.

We make the following formal claim.

Theorem D.1. *There exists a normal-form game, and objective function u_0 of the mediator, such that the unique optimal equilibrium is mixed, and it is impossible to steer players toward that equilibrium using only sublinear payments.*

Proof. Consider the following two-player game. Each player has two actions, A and B. Players 1 and 2 play a coordination game: they score 1 point for playing the same action, and -1 otherwise. The mediator's goal is to *minimize* the welfare of the players.[5]

The welfare-minimizing equilibrium in this game is the fully-mixed one. So, we claim that, using sublinear payments alone, it is impossible to steer players to the mixed equilibrium. Consider the following algorithm for the players: Let $\Gamma^{(t)}$ be the game at time t induced by the mediator's payoff function $p^{(t)}$. Play an arbitrary Nash equilibrium of $\Gamma^{(t)}$, pure if possible. The total regret of the players after T rounds is at most 0 since the players always play a Nash equilibrium. There are three cases:

1. The players play (A, A) or (B, B). In this case, the players get social welfare 2.
2. The players play (A, B) or (B, A). In this case, the players get social welfare -2 in the game itself, but in order for either of these to be a Nash equilibrium, there must be a payment of at least 2 to each player.
3. The players play a mixed strategy. This means that $\Gamma^{(t)}$ had no pure strategy Nash equilibrium. Since (A, A) is not an equilibrium, suppose WLOG that $v_1^{(t)}(\mathsf{B}, \mathsf{A}) > v_1^{(t)}(\mathsf{A}, \mathsf{A})$. Then $p_i^{(t)}(\mathsf{B}, \mathsf{A}) > 1$. Since ($\mathsf{B}$, A) is also not a Nash equilibrium, we have $v_2^{(t)}(\mathsf{B}, \mathsf{B}) > v_2^{(t)}(\mathsf{B}, \mathsf{A})$. Since ($\mathsf{B}$, B) is also not a Nash equilibrium, we have $v_1^{(t)}(\mathsf{A}, \mathsf{B}) > v_1^{(t)}(\mathsf{B}, \mathsf{B})$, so $p_1^{(t)}(\mathsf{A}, \mathsf{B}) > 1$. Thus, all four strategy profiles have either high welfare for the players, or nontrivial payments.

In all three cases, as a result, we must have $\sum_i u_i(\boldsymbol{x}^{(t)}) + 3p_i^{(t)}(\boldsymbol{x}^{(t)}) > 1$ for all timesteps t. Therefore, summing over $t = 1, \dots, T$, it is impossible for both quantities to grow sublinearly in T, which is what would be required for successful steering. □

[4] In the normal-form setting, Monderer & Tennenholtz (2004) also used action recommendations to handle the case of mixed equilibria, but did not discuss the necessity of doing so.

[5] One could construct an example in which the mediator's goal is to *maximize* the players' utility, by simply adding a third player, with one action, whose utility is -10 if P1 and P2 play the same action.

E Omitted Proofs

In this section, we provide the proofs omitted from the main body. Note that no effort was made throughout this paper to optimize the game-dependent or constant factors, so long as they remained polynomial in $|Z|$—they can very likely be improved.

E.1 Proof of Proposition 3.2

We first show Proposition 3.2, the statement of which is recalled below. We highlight that in our following construction we use that players have knowledge of each game.

Proposition 3.2. *There exists a game and some function $R(T) = O(\sqrt{T})$ such that, for all $B \geq 0$, the steering problem is impossible if we add the constraint $\sum_{t=1}^{\infty} \sum_{i=1}^{n} p_i^{(t)}(\boldsymbol{x}^{(t)}) \leq B$.*

Proof. Suppose that the mediator's goal is for the players to coordinate on the equilibrium (B, B) in the coordination 2-player game with the following payoff matrix.

	A	B
A	0.5, 0.5	0,0
B	0,0	1,1

Set $R(T) = 2\sqrt{T}$. We will show that, regardless of the mediator's strategy, it is possible for the players to play (A, A) for all but finitely many rounds.

Suppose the players play as follows. Let $\Gamma^{(t)}$ be the game at time t induced by the mediator's payoff function $p^{(t)}$. For the first B^2 rounds, play an arbitrary Nash equilibrium of $\Gamma^{(t)}$. After that, if (A, A) is a Nash equilibrium of $\Gamma^{(t)}$, play it. Otherwise, play a strategy profile $\boldsymbol{x}^{(t)}$ for which $\sum_{i=1}^{n} p_i^{(t)}(\boldsymbol{x}^{(t)}) > \frac{1}{2}$ (Such a strategy profile must exist, for otherwise (A, A) would be a Nash equilibrium).

The total regret of the players after T rounds is (at most) 0 for $T \leq B^2$, since we have assumed that they are playing a Nash equilibrium of $\Gamma^{(t)}$, and at most $(P+1)k$ for $T > B^2$, where k is the number of times that the final case triggers, since the reward range of $\Gamma^{(t)}$ is at most $[0, P+1]$. But the final case can only trigger at most $2B$ times, since the mediator only has a total budget of B. Therefore, the regret is bounded by $2(P+1)\sqrt{T}/(P+1) = 2\sqrt{T}$ for any T, and for all but $2B + B^2$ rounds, the players are playing a suboptimal equilibrium. So, desideratum (S2) in Definition 3.1 cannot be satisfied. □

E.2 Proof of Theorem 4.2

Theorem 4.2 (Normal-form steering)**.** *Let $p_i(\boldsymbol{x})$ be defined as in (1), set $\alpha = \sqrt{\varepsilon}$, where $\varepsilon := 4nR(T)/T$, and let T be large enough that $\alpha \leq 1$. Then players will be steered toward equilibrium, with both payments and directness gap bounded by $2\sqrt{\varepsilon}$.*

Proof. By construction of the payments, the utility for player i is at least α higher for playing $\boldsymbol{d}_i$ than for any other action, regardless of the actions of the other players. Let $\varepsilon := nR(T)/T$ and $\delta_i^{(t)} := 1 - \boldsymbol{d}_i^\top \boldsymbol{x}_i^{(t)}$. Then the above property ensured by the payments implies that $R(T)/T = \varepsilon/n \geq \alpha \, \mathbb{E}_{t \in [\![T]\!]} \, \delta_i^{(t)}$ Let z^* be the terminal node induced by profile $\boldsymbol{d}$. Then the directness gap is

$$\underset{t}{\mathbb{E}}\left[1 - \hat{\boldsymbol{x}}^{(t)}[z^*]\right] = 1 - \underset{t}{\mathbb{E}} \prod_i (1 - \delta_i^{(t)}) \leq \underset{t}{\mathbb{E}} \sum_i \delta_i^{(t)} \leq \varepsilon/\alpha$$

and the payments are bounded by

$$\underset{t}{\mathbb{E}}\, p_i(\boldsymbol{x}) \leq \alpha + \underset{t}{\mathbb{E}}(1 - \prod_{j \neq i} (1 - \delta_i^{(t)})) \leq \alpha + \varepsilon/\alpha$$

so taking $\alpha = \sqrt{\varepsilon}$ completes the proof. □

E.3 Proof of Theorem 5.2

Before proving Theorem 5.2, we start by showing a useful lemma.

Lemma E.1. *Let $\bar{\boldsymbol{x}}_i := \mathbb{E}_{t\in[\![T]\!]}\, \boldsymbol{x}_i^{(t)}$ for any player $i \in [\![n]\!]$ and $\delta := \sum_{i=1}^{n} \boldsymbol{d}_i^\top(\boldsymbol{d}_i - \bar{\boldsymbol{x}}_i)$. Then, $\mathbb{E}_{t\in[\![T]\!]} \left\|\hat{\boldsymbol{x}}_N^{(t)} - \hat{\boldsymbol{d}}_N\right\|_1 \le |Z|\delta$ for every $N \subseteq [\![n]\!]$. Moreover, if the payments are defined according to (2), the average payment to every player can be bounded by $\mathbb{E}_{t\in[\![T]\!]}\, p_i(\boldsymbol{x}^{(t)}) \le |Z|(2\delta + \alpha)$.*

Proof of Lemma E.1. Let $\delta_i := \boldsymbol{d}_i^\top(\boldsymbol{d}_i - \bar{\boldsymbol{x}}_i)$ for any player $i \in [\![n]\!]$. Then, we have that $\min_{z:\boldsymbol{d}_i[z]=1} \bar{\boldsymbol{x}}_i[z] \ge 1 - \delta_i$, which in turn implies that $\max_{z:\boldsymbol{d}_i[z]=0} \bar{\boldsymbol{x}}_i[z] \le \delta_i$. Now let $N \subseteq [\![n]\!]$. If $z \in Z$ is such that $\boldsymbol{d}_N[z] = 1$,

$$\bar{\boldsymbol{x}}_N[z] = \mathop{\mathbb{E}}_{t\in[\![T]\!]} \boldsymbol{x}_N^{(t)}[z] = \mathop{\mathbb{E}}_{t\in[\![T]\!]} \prod_{j\in N} \boldsymbol{x}_j^{(t)}[z] \ge \mathop{\mathbb{E}}_{t\in[\![T]\!]} \prod_{j\in N} (1 - \delta_j) \ge 1 - \sum_{j\in N} \delta_j = 1 - \delta.$$

Further, if $\boldsymbol{d}_j[z] = 0$ for some $j \in N$,

$$\bar{\boldsymbol{x}}_N[z] \le \bar{\boldsymbol{x}}_j[z] \le \delta_j \le \delta.$$

Thus,

$$\begin{aligned}
\mathop{\mathbb{E}}_{t\in[\![T]\!]} \left\|\hat{\boldsymbol{x}}_N^{(t)} - \hat{\boldsymbol{d}}_N\right\|_1 &= \mathop{\mathbb{E}}_{t\in[\![T]\!]} \left(\sum_{z:\hat{\boldsymbol{d}}_N[z]=0} (\hat{\boldsymbol{x}}_N^{(t)}[z] - \hat{\boldsymbol{d}}_N[z]) + \sum_{z:\hat{\boldsymbol{d}}_N[z]=1} (\hat{\boldsymbol{d}}_N[z] - \hat{\boldsymbol{x}}_N^{(t)}[z]) \right) \\
&= \left\| \mathop{\mathbb{E}}_{t\in[\![T]\!]} \hat{\boldsymbol{x}}_N^{(t)} - \hat{\boldsymbol{d}}_N \right\|_1 = \left\| \bar{\boldsymbol{x}}_N - \hat{\boldsymbol{d}}_N \right\|_1 \le |Z|\delta, \qquad (4)
\end{aligned}$$

since we have shown that $|\bar{\boldsymbol{x}}_N[z] - \hat{\boldsymbol{d}}_N[z]| \le \delta$ for any $z \in Z$. This establishes the first part of the claim. Next, the average payments (2) can by bounded for any player $i \in [\![n]\!]$ as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{t\in[\![T]\!]} \Bigg[\left[u_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{d}_{-i}) - u_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{x}_{-i}^{(t)}) \right] \\
&\qquad - \min_{\boldsymbol{x}_i' \in X_i} \left[u_i(\boldsymbol{x}_i', \boldsymbol{d}_{-i}) - u_i(\boldsymbol{x}_i', \boldsymbol{x}_{-i}^{(t)}) \right] + \alpha \boldsymbol{d}_i^\top \boldsymbol{x}_i^{(t)} \Bigg] \\
&\qquad\qquad \le 2 \mathop{\mathbb{E}}_{t\in[\![T]\!]} \left\| \hat{\boldsymbol{x}}_{-i}^{(t)} - \hat{\boldsymbol{d}}_{-i} \right\|_1 + \alpha|Z| \le |Z|(2\delta + \alpha),
\end{aligned}$$

where we used the normalization assumption $|u_i(\cdot)| \le 1$, and the fact that $\boldsymbol{d}_i^\top \boldsymbol{x}_i^{(t)} \le |Z|$. This concludes the proof. □

We are now ready to prove Theorem 5.2, the formal version of which is recalled below.

Theorem 5.2. *Set $\alpha = \sqrt{\varepsilon}$, where $\varepsilon := 4nR(T)/T$, and let T be large enough that $\alpha \le 1/|Z|$. Then,* FullFeedbackSteer *results in average realized payments and directness gap at most $3|Z|\sqrt{\varepsilon}$.*

Proof. The utility of each player $i \in [\![n]\!]$ reads

$$v_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i}) := \alpha \boldsymbol{d}_i^\top \boldsymbol{x}_i + u_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i}) - \min_{\boldsymbol{x}_i' \in X_i} [u_i(\boldsymbol{x}_i', \boldsymbol{d}_{-i}) - u_i(\boldsymbol{x}_i', \boldsymbol{x}_{-i})].$$

Given that $\boldsymbol{d}$ is an equilibrium, it follows that $\boldsymbol{d}_i$ is a strict best response for any player $i \in [\![n]\!]$. That is, the regret of each player $i \in [\![n]\!]$ after T iterations can be lower bounded as

$$\sum_{t=1}^{T} \left(\alpha \boldsymbol{d}_i^\top(\boldsymbol{d}_i - \boldsymbol{x}_i^{(t)}) + u_i(\boldsymbol{d}) - u_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{d}_{-i}) \right) \ge \alpha T \boldsymbol{d}_i^\top(\boldsymbol{d}_i - \bar{\boldsymbol{x}}_i),$$

where we used that $u_i(\boldsymbol{d}) - u_i(\boldsymbol{x}_i^{(t)}, \boldsymbol{d}_{-i}) \ge 0$ since $\boldsymbol{d}$ is an equilibrium. Thus,

$$\sum_{i=1}^{n} \boldsymbol{d}_i^\top(\boldsymbol{d}_i - \bar{\boldsymbol{x}}_i) \le \frac{nR(T)}{\alpha T} = \frac{\varepsilon}{\alpha}.$$

We can now apply Lemma E.1 to obtain that $\mathbb{E}_{t \in [\![T]\!]} \left\| \hat{\boldsymbol{x}}^{(t)} - \hat{\boldsymbol{d}} \right\|_1 \le |Z|\delta$, where $\delta := \varepsilon/\alpha$. Thus, the directness gap is bounded by

$$\left| \mathop{\mathbb{E}}_{t \in [\![T]\!]} \mathop{\mathbb{E}}_{z \sim \boldsymbol{x}^{(t)}} (1 - \hat{\boldsymbol{d}}[z]) \right| \le \left\| \mathop{\mathbb{E}}_{t} \hat{\boldsymbol{x}}^{(t)} - \hat{\boldsymbol{d}} \right\|_1 = \mathop{\mathbb{E}}_{t} \left\| \hat{\boldsymbol{x}}^{(t)} - \hat{\boldsymbol{d}} \right\|_1 \le \frac{n|Z|R(T)}{\alpha T},$$

where the first inequality uses that $|u_0(\cdot)| \le 1$, and the equality follows because $\hat{\boldsymbol{d}}$ is an extreme point of X (as in (4)). Furthermore, by Lemma E.1, the payment to each player $i \in [\![n]\!]$ can be bounded by

$$2|Z|(2\delta + \alpha) = 2|Z|\frac{\varepsilon}{\alpha} + |Z|\alpha.$$

As a result, setting $\alpha = \sqrt{\varepsilon}$ for T sufficiently large so that $\alpha \le 1/|Z|$, we guarantee that the payment to each player is bounded by $3n|Z|\sqrt{\varepsilon}$ and the directness gap is bounded by $|Z|\sqrt{\varepsilon}$, as desired. □

E.4 Proof of Theorem 6.5

We next provide the proof of Theorem 6.5.

Theorem 6.5. *Set the hyperparameters $\alpha = \varepsilon^{2/3}|Z|^{-1/3}$ and $\lambda = |Z|^{2/3}\varepsilon^{-1/3}$, where $\varepsilon := (R_0(T) + 4nR(T))/T$ is the average regret bound summed across players, and let T be large enough that $\alpha \le 1/|Z|$. Then running* OnlineSteer *results in average realized payments, directness gap, and optimality gap all bounded by $7\lambda^*|Z|^{4/3}\varepsilon^{1/3}$.*

Proof. To simplify the notation, we assume without loss of generality that $u_0^* = 0$. We will also use the change of variables $\boldsymbol{y} := \boldsymbol{x} - \boldsymbol{d} \in Y := X - \boldsymbol{d}$. With the payments and utility functions as specified, the losses given to the players and the mediator are, up to additive constants, exactly the losses that they would see if they were playing the zero-sum game

$$\max_{\boldsymbol{\mu} \in \Xi} \min_{\boldsymbol{y} \in Y} \frac{1}{\lambda} \boldsymbol{c}^\top \boldsymbol{\mu} - \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{y} - \alpha \boldsymbol{d}^\top \boldsymbol{y}, \tag{5}$$

where $\mathbf{A} = [\mathbf{A}_1 \quad \cdots \quad \mathbf{A}_n]$, $\boldsymbol{d} = \begin{bmatrix} \boldsymbol{d}_1^\top & \cdots & \boldsymbol{d}_n^\top \end{bmatrix}^\top$, and $\lambda \ge 1$. Now let $(\lambda^*, \boldsymbol{y}^*)$ be an optimal dual solution in (3). If we select $\lambda \ge \lambda^*$ and $\boldsymbol{y}' := (\lambda^*/\lambda)\boldsymbol{y}^*$, then $(\lambda, \boldsymbol{y}')$ is also an optimal dual solution in (3). Therefore,

$$\max_{\boldsymbol{\mu} \in \Xi} \min_{\boldsymbol{y} \in Y} \frac{1}{\lambda} \boldsymbol{c}^\top \boldsymbol{\mu} - \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{y} - \alpha \boldsymbol{d}^\top \boldsymbol{y} \le -\alpha \boldsymbol{d}^\top \boldsymbol{y}',$$

since it is assumed that $u_0^* = 0$. Further, we know that $(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{y}})$ is an ε-Nash equilibrium of the above zero-sum game since $(R_0(T) + 4nR(T))/T = \varepsilon$; in particular, we have that[6]

$$-\alpha \boldsymbol{d}^\top \bar{\boldsymbol{y}} \le \max_{\boldsymbol{\mu} \in \Xi} \frac{1}{\lambda} \boldsymbol{c}^\top \boldsymbol{\mu} - \boldsymbol{\mu}^\top \mathbf{A} \bar{\boldsymbol{y}} - \alpha \boldsymbol{d}^\top \bar{\boldsymbol{y}} \le -\alpha \boldsymbol{d}^\top \boldsymbol{y}' + \varepsilon,$$

or, rearranging,

$$-\boldsymbol{d}^\top \bar{\boldsymbol{y}} \le -\frac{\lambda^*}{\lambda} \boldsymbol{d}^\top \boldsymbol{y}^* + \frac{\varepsilon}{\alpha} \le \frac{\lambda^*}{\lambda}|Z| + \frac{\varepsilon}{\alpha} := \delta.$$

Thus, by Lemma E.1, the average payment is bounded by $|Z|(2\delta + \alpha)$. We now turn to the mediator's average utility. The equilibrium value of (5) is at least $-\alpha|Z|$ (achieved by the optimal equilibrium), in turn implying that the current value in the game under $(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{y}})$ is at least $-\alpha|Z| - \varepsilon$. So,

$$\mathop{\mathbb{E}}_{t \in [\![T]\!]} u_0(\boldsymbol{\mu}^{(t)}, \boldsymbol{d}) = \boldsymbol{c}^\top \bar{\boldsymbol{\mu}} \ge \min_{\boldsymbol{y} \in Y} \boldsymbol{c}^\top \bar{\boldsymbol{\mu}} - \lambda \left[\bar{\boldsymbol{\mu}}^\top \mathbf{A} \boldsymbol{y} - \alpha \boldsymbol{d}^\top \boldsymbol{y} \right] \ge -\lambda(\alpha|Z| + 2\varepsilon).$$

By Lemma E.1 again,

$$\left| \mathop{\mathbb{E}}_{t \in [\![T]\!]} u_0(\boldsymbol{\mu}^{(t)}, \boldsymbol{x}^{(t)}) - u_0(\boldsymbol{\mu}^{(t)}, \boldsymbol{d}) \right| \le \left\| \mathop{\mathbb{E}}_{t \in [\![T]\!]} \hat{\boldsymbol{x}}^{(t)} - \hat{\boldsymbol{d}} \right\|_1 \le \mathop{\mathbb{E}}_{t \in [\![T]\!]} \left\| \hat{\boldsymbol{x}}^{(t)} - \hat{\boldsymbol{d}} \right\|_1 \le |Z|\delta,$$

[6] A technical comment here: $-\boldsymbol{d}^\top \boldsymbol{y}$ is *nonnegative*, and takes its *minimum* value at $\boldsymbol{y} = 0$.

so the optimality gap is bounded by $2\varepsilon\lambda + |Z|\alpha\lambda + |Z|\delta$, and the directness gap is bounded by $|Z|\delta$. It thus suffices to select hyperparameters α and λ so as to minimize the following expression, which is an upper bound on all three gaps:

$$2\varepsilon\lambda + |Z|\alpha\lambda + 2|Z|\delta = 2\varepsilon\lambda + |Z|\alpha\lambda + 2|Z|^2\frac{\lambda^*}{\lambda} + 2|Z|\frac{\varepsilon}{\alpha}.$$

In particular, setting the hyperparameters as in the theorem statement and plugging them into the expression above, we arrive at the bound

$$2\varepsilon^{2/3}|Z|^{2/3} + |Z|^{4/3}\varepsilon^{1/3} + 2\lambda^*|Z|^{4/3}\varepsilon^{1/3} + 2|Z|^{4/3}\varepsilon^{1/3} \le 7\lambda^*|Z|^{4/3}\varepsilon^{1/3},$$

as claimed. □

It is worth noting that, despite the fact that it would speed up the convergence, we cannot set λ and α dependent on λ^*, because we do not know λ^* *a priori*.

E.5 Proof of Theorem 5.4

We continue with the proof of Theorem 5.4.

Theorem 5.4. *For every $P > 0$, there exists an extensive-form game Γ with $O(P)$ players, $O(P^2)$ nodes, and rewards bounded in $[0, 1]$ such that, with payments $q_i^{(t)} : Z \to [0, P]$, it is impossible to steer players to the welfare-maximizing Nash equilibrium, even when $R(T) = 0$.*

Proof. For any $n > 0$, consider the following n-player extensive-form game Γ, which has $O(n^2)$ nodes. Every player has only a single information set with two actions, and we will (for good reason, as we will see later) refer to the actions as Stag and Hare. Chance first picks some $j \in [\![n]\!] \cup \{\perp\}$ uniformly at random.

If $j \neq \perp$, then player j plays an action (which is either Stag or Hare). If i plays Hare, it gets utility $1/2$; otherwise, it gets utility 0. All other players get utility 0.

If $k = \perp$, chance samples another player k uniformly at random from $[\![n]\!]$. Then, in the order $k, k+1, \dots, n, 1, 2, \dots, k-1$, the players play their actions. If any player at any point plays Hare, then the game ends and all players get 0. If all players play Stag, then all players get 1.

The normal form of this game is an n-player generalization of the Stag Hunt game: if all players play Stag then all players have (expected) payoff $1/(n+1)$; if any player plays Hare then every player has expected payoff $(1/2)/(n+1)$ for playing Hare and 0 for playing Stag. In particular, the welfare-optimal profile, "everyone plays Stag", is a Nash equilibrium, and hence is also the welfare-optimal EFCE, with social welfare $n/(n+1)$. "Everyone plays Hare" is also an equilibrium, with social welfare $(1/2)n/(n+1)$. The game tree when $n = 3$ is depicted in Figure 2.

Intuitively, the rest of the proof works as follows. Suppose that all players are currently playing Hare. The mediator needs to incentivize players to play Stag, but it has a dilemma. It cannot give a large payment to i for playing Stag when $j = i$—then the average payment for each player would diverge if the players were to move to the Stag equilibrium. The only other location that the mediator could possibly give a payment to i is when $j = \perp$, $k = i$, player i plays Stag, and the next player plays Hare. But this node is only reached with probability $O(1/n^2)$—therefore, to outweigh i's current incentive of $\Theta(1/n)$ of playing Hare, the payment at this node would have to be $\Theta(n)$, at which point taking $n = \Theta(P)$ would complete the proof.

We now formalize this intuition. Take $n = \lceil 4P \rceil$. Consider players who play as follows. At each timestep t, the players consider the extensive-form game $\Gamma^{(t)}$ induced by adding the payment functions $q_i^{(t)}$ that the mediator would play, and ignoring mediator recommendations. That is, $\Gamma^{(t)}$ is identical to Γ except that $q_i^{(t)}$ has been added to player i's utility function. If "everyone plays Hare" is a Nash equilibrium of $\Gamma^{(t)}$, all players play Hare. Otherwise, the players play according to an arbitrary Nash equilibrium of $\Gamma^{(t)}$.

Since the players are playing according to a Nash equilibrium at every step, they all have regret at most 0. Now consider two cases.

1. There is a player i such that plays Stag with probability less than $1/2$. Then the social welfare is at most $(3/4)n/(n+1)$, which is lower than the optimal social welfare by $(1/4)n/(n+1)$.

2. All players play Stag with probability at least $1/2$. Then, in particular, "everyone plays Hare" is not a Nash equilibrium in $\Gamma^{(t)}$. So, if everyone were to play Hare, there is some player i who would rather deviate and play Stag. Thus, the mediator must be giving an expected payment to i of at least $(1/2)/(n+1)$. As discussed above, there are only two nodes z for which the setting of $q_i^{(t)}(z)$ increases i's utility for playing Stag relative to its utility for playing Hare. The first is when $j = \perp$, $k = i$, i plays Stag, and the next player plays Hare. Since $P \le n/4$ and this node occurs with probability $1/(n(n+1))$, even the maximum payment at this node contributes at most $(1/4)/(n+1)$ to the expected payment. Therefore, the remainder of the payment, $(1/2)/(n+1)$, must be given when $j = i$ and then i plays Stag. But Player i plays Stag with probability at least $1/2$, so i's observed expected payment is at least $(1/4)/(n+1)$.

Therefore, we have

$$\left(u_0^* - \mathbb{E}\, u_0(z^{(t)})\right) + \mathbb{E} \sum_{i \in [\![n]\!]} q_i^{(t)}(z^{(t)}) \ge \frac{1}{4(n+1)}$$

where u_0 is the social welfare function, so it is impossible for both quantities to tend to 0 as $T \to \infty$. □

E.6 Proof of Theorem 5.6

Finally, we conclude with the proof of Theorem 5.6.

Theorem 5.6. *Set the hyperparameters $\alpha = 4|Z|^{1/2}\varepsilon^{1/4}$ and $P = 2|Z|^{1/2}\varepsilon^{-1/4}$, where $\varepsilon := R(T)/T$, and let T be large enough that $\alpha \le 1$. Then, running* BanditSteer *for T rounds results in average realized payments bounded by $8|Z|^{1/2}\varepsilon^{1/4}$, and directness gap by $2\varepsilon^{1/2}$.*

We use the following notation.

- The set D_S is the set of nodes at which all players in set S have played directly: $D_S = \{z \in Z : \boldsymbol{d}_i[z] = 1 \forall i \in S\}$. The set $D'_S = Z \setminus D_S$ is its complement.

- $\boldsymbol{x}$ is a random variable for the correlated strategy profile played by all players through the T timesteps. That is, $\boldsymbol{x}$ is a uniform sample from $\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(T)}\}$.

- $\pi(S|\boldsymbol{x})$ is the probability that a terminal node from set S is reached, given that the mediator plays $\boldsymbol{\mu}$ and the players play the (possibly correlated) strategy profile $\boldsymbol{x}$. That is, $\pi(S|\boldsymbol{x}) = \Pr_{z \sim (\boldsymbol{\mu}, \boldsymbol{x})}[z \in S]$.

- $\tilde{u}_i(\boldsymbol{x}) = u_i(\boldsymbol{x}) + \mathbb{E}_{z \sim (\boldsymbol{\mu}, \boldsymbol{x})}\, q_i(z)$ is the expected utility for player i, including payment, under profile $(\boldsymbol{\mu}, \boldsymbol{x})$.

- $u_i(\boldsymbol{y}_i - \boldsymbol{x}_i, \boldsymbol{x}_{-i}) := u_i(\boldsymbol{y}_i, \boldsymbol{x}_{-i}) - u_i(\boldsymbol{x}_i, \boldsymbol{x}_{-i})$ is player i's advantage for playing $\boldsymbol{y}_i$ instead of $\boldsymbol{x}$. $\tilde{u}_i(\boldsymbol{y}_i - \boldsymbol{x}_i, \boldsymbol{x}_{-i})$ and $\pi(z|\boldsymbol{y}_i - \boldsymbol{x}_i, \boldsymbol{x}_{-i})$ are defined similarly.

Let $\varepsilon = R(T)/T$. Then after T timesteps, since the players are no-regret learners, their average joint strategy profile will be an $P\varepsilon$-NFCCE of the extensive-form game with the payments added.

Intuitively, the proof will go as follows. We will show that, for P sufficiently large, each player's incentive to be direct will be *at least* as great as it would have been if everyone else were also direct, plus α. Then it will follow from the fact that $\boldsymbol{\mu}$ is an equilibrium, and picking $\alpha \gg P\varepsilon$, that all players must therefore be direct. We first prove a lemma. Informally, the lemma states that, when any player i deviates, all other players must be direct.

Lemma E.2. *Let z be any node with $\boldsymbol{d}_i[z] = 0$, that is, any node at which player i has deviated. Then $|\pi(z|\boldsymbol{x}_i, \boldsymbol{d}_{-i} - \boldsymbol{x}_{-i})| \le \gamma := n\varepsilon + \sum_j \delta_j/P$, where $\delta_j := u_j(\boldsymbol{x}_j - \boldsymbol{d}_j, \boldsymbol{x}_{-j})$ is player j's current deviation benefit.*

Proof. Assume without loss of generality that $i = 1$, and consider two cases.

1. $\boldsymbol{d}_j[z] = 0$ for some $j \neq i$—that is, some other player has also deviated. Then $\pi(z|\boldsymbol{x}_i, \boldsymbol{d}_{-i}) = 0$. Assume for contradiction that $\pi(z|\boldsymbol{x}) > \gamma$. Let $h_i, h_j \prec z$ be the two deviation points—that is, $\boldsymbol{d}_i[h_i] = 1$ but $\boldsymbol{d}_i[h_i a_i] = 0$ where $h_i a_i \preceq z$, and similar for h_j. Suppose without loss that $h_i \prec h_j$. Now consider player j's incentive. If player j were to switch to playing $\boldsymbol{d}_j$, its expected payment increases by at least γP, and its expected utility (sans payment) decreases by δ_j, by definition. When $\gamma \ge \varepsilon + \delta_j/P$, this produces a contradiction.

2. $\boldsymbol{d}_j[z] = 1$ for all $j \neq i$. Then $\pi(z|\boldsymbol{x}_i, \boldsymbol{d}_{-i}) \ge \pi(z|\boldsymbol{x})$, so we need to show that $\pi(z|\boldsymbol{x}_i, \boldsymbol{d}_{-i}) - \pi(z|\boldsymbol{x}^{(t)}) \le \gamma$. That is, other players will almost always play to catch player i deviating, whenever possible. Suppose not. Let $h \prec z$ be the point where player i deviated (that is, $\boldsymbol{d}_i[h] = 1$ but $\boldsymbol{d}_i[ha_1] = 0$ where $ha_1 \preceq z$). Let a_0 be the direct action at h. Notice that, for any player $j \neq i$, if j shifts to playing the direct strategy, the probability of leaving the path to ha before reaching ha itself cannot increase by more than $\varepsilon + \delta_i/P$: otherwise, player j's expected utility would be increasing by more than δ_i, a contradiction. If all $n-1$ players allocate their deviations in this manner, and even if the remaining $(n-1)\delta_i/P$ probability of leaving path ha is then all allocated to node z, the reach probability of z could not have increased by more than $\sum_j(\varepsilon + \delta_j/P)$. Thus, when γ is larger than this value, we have a contradiction. □

The rest of the proof is structured as follows. We will first show, roughly speaking, that *player i's deviation benefit*—that is, its advantage for playing $\boldsymbol{x}_i^{(t)}$ at each timestep t instead of playing $\boldsymbol{d}_i$—is *smaller* against the opponent strategies $\boldsymbol{x}_{-i}^{(t)}$ than it would be against $\boldsymbol{d}_{-i}^{(t)}$, modulo a small additive error. Then, the proof will follow from the fact that $\boldsymbol{d}$ is an equilibrium against $\boldsymbol{\mu}$, so therefore all players should play according to $\boldsymbol{d}$.

$$\begin{aligned}
&\tilde{u}_i(\boldsymbol{x}) - \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i}) \\
&= \sum_{z \in D_i \cap D_{-i}} \tilde{u}_i(z)[\pi(z|\boldsymbol{x}) - \pi(z|\boldsymbol{x}_i, \boldsymbol{d}_{-i})] + \sum_{z \in D_i \cap D'_{-i}} \tilde{u}_i(z)\pi(z|\boldsymbol{x}) \\
&\quad + \underbrace{\sum_{z \in D'_i} u_i(z)[\pi(z|\boldsymbol{x}) - \pi(z|\boldsymbol{x}_i, \boldsymbol{d}_{-i})]}_{\le \gamma|Z|} \\
&\le \sum_{z \in D_i \cap D_{-i}} \tilde{u}_i(z)[\pi(z|\boldsymbol{x}) - \pi(z|\boldsymbol{x}_i, \boldsymbol{d}_{-i})] + \sum_{z \in D_i \cap D'_{-i}} \tilde{u}_i(z)\pi(z|\boldsymbol{x}) + \gamma|Z|
\end{aligned}$$

where we use, in order, the definition of expected utility, the fact that $u_i(z) = \tilde{u}_i(z)$ when $\boldsymbol{d}_i[z] = 0$ and $\pi(z|\boldsymbol{x}) = 0$ whenever $\boldsymbol{x}_i[z] = 0$ for any i, and finally Lemma E.2. Similarly,

$$\begin{aligned}
&\tilde{u}_i(\boldsymbol{d}_i, \boldsymbol{x}_{-i}) - \tilde{u}_i(\boldsymbol{d}) \\
&= \sum_{z \in D_i \cap D_{-i}} \tilde{u}_i(z)[\pi(z|\boldsymbol{d}_i, \boldsymbol{x}_{-i}) - \pi(z|\boldsymbol{d})] + \sum_{z \in D_i \cap D'_{-i}} \tilde{u}_i(z)\pi(z|\boldsymbol{d}_i, \boldsymbol{x}_{-i}).
\end{aligned}$$

Thus,

$$\begin{aligned}
&P\varepsilon - [\tilde{u}_i(\boldsymbol{d}) - \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i})] \\
&\ge [\tilde{u}_i(\boldsymbol{d}_i, \boldsymbol{x}_{-i}) - \tilde{u}_i(\boldsymbol{x})] - [\tilde{u}_i(\boldsymbol{d}) - \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i})] \\
&\ge \sum_{z \in D_i \cap D_{-i}} \tilde{u}_i(z) \underbrace{[\pi(z|\boldsymbol{d}_i - \boldsymbol{x}_i, \boldsymbol{x}_{-i}) - \pi(z|\boldsymbol{d}_i - \boldsymbol{x}_i, \boldsymbol{d}_{-i})]}_{\le 0}
\end{aligned}$$

$$
\begin{aligned}
&+ 2 \sum_{z \in D_i \cap D'_{-i}} \pi(z|\boldsymbol{d}_i - \boldsymbol{x}_i, \boldsymbol{x}_{-i}) - \gamma|Z| \\
&\geq 2 \sum_{z \in D_i \cap D_{-i}} [\pi(z|\boldsymbol{d}_i - \boldsymbol{x}_i, \boldsymbol{x}_{-i}) - \pi(z|\boldsymbol{d}_i - \boldsymbol{x}_i, \boldsymbol{d}_{-i})] \\
&\quad + 2 \sum_{z \in D_i \cap D'_{-i}} \pi(z|\boldsymbol{d}_i - \boldsymbol{x}_i, \boldsymbol{x}_{-i}) - \gamma|Z| \\
&= 2[\pi(D'_i|\boldsymbol{x}_i, \boldsymbol{d}_{-i}) - \pi(D'_i|\boldsymbol{x})] - \gamma|Z| \geq -3\gamma|Z|.
\end{aligned}
$$

The first inequality uses the fact $\pi(z|\boldsymbol{d}_i, \boldsymbol{x}_{-i}) - \pi(z|\boldsymbol{x}) \geq 0$ when $\boldsymbol{d}_i[z] = 1$ and $\tilde{u}_i(z) \geq P \geq 2$ when $\boldsymbol{d}_i[z] = 1$ and $\boldsymbol{d}_{-i}[z] = 0$. The quantity in braces is nonpositive because for any profile $\boldsymbol{x}$, setting $\boldsymbol{x}_{-i} = \boldsymbol{d}$ only increases the probability that player i is the one to deviate from the path to z. The second inequality uses the nonpositivity of the quantity in the braces, and the fact that $\tilde{u}_i(z) = u_i(z) + \alpha \leq 2$.

Now we look at the remaining quantity, $\tilde{u}_i(\boldsymbol{d}) - \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i})$, which is simply the negative deviation of benefit of Player i's strategy $\boldsymbol{x}_i$ if all other players were direct. Indeed, since we know that $\boldsymbol{\mu}$ is an equilibrium, we have

$$
\begin{aligned}
&\tilde{u}_i(\boldsymbol{d}) - \tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i}) \\
&= \underbrace{[\tilde{u}_i(\boldsymbol{d}) - u_i(\boldsymbol{d})]}_{=\alpha} - \underbrace{[\tilde{u}_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i}) - u_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i})]}_{=\alpha(1-\Delta_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i}))} + \underbrace{[u_i(\boldsymbol{d}) - u_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i})]}_{\geq 0} \\
&\geq \alpha\Delta_i(\boldsymbol{x}_i, \boldsymbol{d}_{-i}) \geq \alpha\Delta_i(\boldsymbol{x}) - \gamma|Z|,
\end{aligned}
$$

where the final inequality again uses Lemma E.2 and $\Delta_i(\boldsymbol{x}) := \sum_{z:\boldsymbol{d}_i[z]=0} \pi(z|\boldsymbol{x})$

Now, notice that $\delta_i \leq \Delta_i(\boldsymbol{x})$, by definition. Substituting into the previous inequality and Lemma E.2, we have

$$
\alpha\Delta_i(\boldsymbol{x}) - 4\left(n\varepsilon + \frac{\sum_j \Delta_j(\boldsymbol{x})}{P}\right)|Z| \leq P\varepsilon,
$$

or, rearranged,

$$
\alpha\Delta_i(\boldsymbol{x}) - 4|Z|\frac{\sum_j \Delta_j(\boldsymbol{x})}{P} \leq (P + 4n)\varepsilon \leq 2P\varepsilon
$$

when $P \geq 4n$. Summing over all players i yields

$$
\alpha\Delta - 4|Z|\frac{\Delta}{P} \leq (P + 4n)\varepsilon \leq 2P\varepsilon
$$

where $\Delta = \sum_i \Delta_i(\boldsymbol{x})$, or, rearranging,

$$
\Delta \leq \frac{2P\varepsilon}{\alpha - 4|Z|/P}.
$$

Both the payments from the mediator and the gap to optimal value are thus bounded by

$$
\alpha + P\Delta \leq \alpha + \frac{2P^2\varepsilon}{\alpha - 4|Z|/P}.
$$

Now taking $\alpha = 4|Z|^{1/2}\varepsilon^{1/4}$ and $P = 2|Z|^{1/2}/\varepsilon^{1/4}$ gives the desired bounds.

F Bandit Online Steering in Normal-Form Games

Essentially, the algorithm replicates the OnlineSteer algorithm (Algorithm 6.4) by randomly sampling. In the normal-form setting, a mediator pure strategy is a profile of actions, $d^{(t)} \in A_1 \times \cdots \times A_n$, where A_i is the action set of player i. Each player i observes the recommendation $d_i^{(t)}$, and chooses an action $a_i^{(t)}$.

Definition F.1 (NormalFormSteer)**.** The mediator runs a *bandit* regret minimization algorithm $\mathcal{R}_0$, such as Exp3 (Auer et al., 2002), over its own strategy space X_0, which we assume has regret at most $R_0(T)$ after T rounds. On each round, the mediator does the following.

1. Get a strategy $d^{(t)} = (d_1^{(t)}, \dots, d_n^{(t)})$ from $\mathcal{R}_0$.
2. With probability α, ignore $d^{(t)}$ and recommend actions $(\tilde{d}_1^{(t)}, \dots, \tilde{d}_n^{(t)})$ uniformly at random. Let $a^{(t)} = (a_1^{(t)}, \dots, a_n^{(t)})$ be the tuple of actions played by the players. Pay each player
$$q_i^{(t)}(a^{(t)}) := 1 - u_i(a^{(t)}) + \mathbb{1}\left\{a_i^{(t)} = \tilde{d}_i^{(t)}\right\}.$$
Pass reward 0 to the mediator.
3. Otherwise, give recommendation $r_i^{(t)}$ to each player i. Pay each player
$$q_i^{(t)}(a^{(t)}) := u_i(a_i^{(t)}, d_{-i}^{(t)}) - u_i(a^{(t)}) - \min_{a_i' \in A_i} \left[u_i(a_i', d_{-i}^{(t)}) - u_i(a_i', a_{-i}^{(t)})\right].$$
Pass reward $\frac{1}{\lambda} u_0(d) - \sum_{i=1}^n \left[u_i(a_i^{(t)}, d_{-i}) - u_i(d)\right]$ to the mediator.

Theorem F.2. *Set the hyperparameters $\alpha = |Z|^{1/3} n^{-2/3} b^{1/3} \varepsilon^{2/3}$ and $\lambda = |Z|^{1/3} n^{1/3} b^{1/3} \varepsilon^{-1/3}$ where $\varepsilon := (R_0(T) + 4nR(T))/T$ is the average regret bound summed across players and $b = \max_i |A_i|$. Let T be large enough that $\alpha \le 1/(2n)$. Then running* NormalFormSteer *results in average realized payments, directness gap, and optimality gap all bounded by $10\lambda^* |Z|^{4/3} \varepsilon^{1/3}$.*

Proof. Reverting to the extensive-form notation, the expected utility of the mediator on iteration t is
$$(1 - \alpha)\left(\frac{1}{\lambda} u_0(\boldsymbol{\mu}^{(t)}, \boldsymbol{d}) - \sum_{i=1}^n \left[u_i(\boldsymbol{\mu}^{(t)}, \boldsymbol{x}_i^{(t)}, \boldsymbol{d}_{-i}) - u_i(\boldsymbol{\mu}^{(t)}, \boldsymbol{d})\right]\right)$$
The expected utility of player i is, up to an additive term that cannot be affected by player i,
$$\alpha \frac{1}{|A_i|} \boldsymbol{d}_i^\top \boldsymbol{x}_i + (1 - \alpha)\Big(u_i(\boldsymbol{\mu}^{(t)}, \boldsymbol{x}^{(t)}, \boldsymbol{d}_{-i}) - u_i(\boldsymbol{\mu}^{(t)}, \boldsymbol{d}^{(t)}, \boldsymbol{d}_{-i})\Big)$$
Therefore, the players and mediator experience the same utilities that they would in the zero-sum game
$$\max_{\boldsymbol{\mu} \in \Xi} \min_{\boldsymbol{y} \in Y} \; (1 - \alpha)\left(\frac{1}{\lambda} \boldsymbol{c}^\top \boldsymbol{\mu} - \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{y}\right) - \alpha \sum_i \frac{1}{|A_i|} \boldsymbol{d}_i^\top \boldsymbol{y}_i, \tag{6}$$
where, as in the proof of Theorem 6.5, $\boldsymbol{y} := \boldsymbol{x} - \boldsymbol{d}$. Following the proof of Theorem 6.5, we conclude that $(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{y}})$ must be an ε-Nash equilibrium of the above zero-sum game. Let $\lambda^*, \lambda, \boldsymbol{y}^*, \boldsymbol{y}'$ be as in that proof. For simplicity of notation, let $\boldsymbol{D}$ be the vector satisfying $\boldsymbol{D}^\top \boldsymbol{y} = \sum_i \frac{1}{|A_i|} \boldsymbol{d}_i^\top \boldsymbol{y}_i$, Then
$$-\alpha \boldsymbol{D}^\top \bar{\boldsymbol{y}} \le \max_{\boldsymbol{\mu} \in \Xi} \min_{\boldsymbol{y} \in Y} \; (1 - \alpha)\left(\frac{1}{\lambda} \boldsymbol{c}^\top \boldsymbol{\mu} - \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{y}\right) - \alpha \boldsymbol{D}^\top \boldsymbol{y} \le -\alpha \boldsymbol{D}^\top \boldsymbol{y}' + \varepsilon$$
or, rearranging,[7]
$$-\frac{1}{b} \boldsymbol{d}^\top \bar{\boldsymbol{y}} \le -\boldsymbol{D}^\top \bar{\boldsymbol{y}} \le -\frac{\lambda^*}{\lambda} \boldsymbol{D}^\top \boldsymbol{y}^* + \frac{\varepsilon}{\alpha} \le \frac{\lambda^*}{\lambda} n + \frac{\varepsilon}{\alpha} := \frac{\delta}{b}.$$
where $b = \max_i |A_i|$ is the maximum branching factor. Thus, by Lemma E.1, the average payment is bounded by $|Z|(2\delta + \alpha)$. We now turn to the mediator's average utility. The equilibrium value of (6) is at least $-\alpha n$ (achieved by the optimal equilibrium), in turn implying that the current value in the game under $(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{y}})$ is at least $-\alpha n - \varepsilon$. So,
$$\mathop{\mathbb{E}}_{t \in [\![T]\!]} u_0(\boldsymbol{\mu}^{(t)}, \boldsymbol{d}) = \boldsymbol{c}^\top \bar{\boldsymbol{\mu}} \ge \min_{\boldsymbol{y} \in Y} \; \boldsymbol{c}^\top \bar{\boldsymbol{\mu}} - \lambda \left[\bar{\boldsymbol{\mu}}^\top \mathbf{A} \boldsymbol{y} - \frac{\alpha}{1 - \alpha} \boldsymbol{d}^\top \boldsymbol{y}\right] \ge -2\lambda(\alpha n + 2\varepsilon).$$

[7] We note once again that $-\boldsymbol{d}^\top \bar{\boldsymbol{y}}$ and $-\boldsymbol{D}^\top \bar{\boldsymbol{y}}$ are, despite the negative sign, a *nonnegative* quantities since $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{d}$.

since $\alpha \le 1/2$. By Lemma E.1 again,

$$\left| \mathop{\mathbb{E}}_{t \in [\![T]\!]} u_0(\boldsymbol{\mu}^{(t)}, \boldsymbol{x}^{(t)}) - u_0(\boldsymbol{\mu}^{(t)}, \boldsymbol{d}) \right| \le \left\| \mathop{\mathbb{E}}_{t \in [\![T]\!]} \hat{\boldsymbol{x}}^{(t)} - \hat{\boldsymbol{d}} \right\|_1 \le \mathop{\mathbb{E}}_{t \in [\![T]\!]} \left\| \hat{\boldsymbol{x}}^{(t)} - \hat{\boldsymbol{d}} \right\|_1 \le |Z|\delta,$$

so the optimality gap is bounded by $4\varepsilon\lambda + 2n\alpha\lambda + |Z|\delta$, and the directness gap is bounded by $|Z|\delta$. It thus suffices to select hyperparameters α and λ so as to minimize the following expression, which is an upper bound on all three gaps:

$$4\varepsilon\lambda + 2n\alpha\lambda + 2|Z|\delta \le 4\varepsilon\lambda + 2n\alpha\lambda + 2|Z|nb\frac{\lambda^*}{\lambda} + 2|Z|b\frac{\varepsilon}{\alpha}.$$

In particular, setting the hyperparameters

$$\alpha = |Z|^{1/3} n^{-2/3} b^{1/3} \varepsilon^{2/3} \quad \text{and} \quad \lambda = |Z|^{1/3} n^{1/3} b^{1/3} \varepsilon^{-1/3}$$

we arrive at the bound

$$4\varepsilon^{2/3}(|Z|nb)^{1/3} + 2(|Z|nb)^{2/3}\varepsilon^{-1/3} + 2\lambda^*(|Z|nb)^{2/3}\varepsilon^{-1/3} + 2(|Z|nb)^{2/3}\varepsilon^{-1/3} \le 10\lambda^*|Z|^{4/3}\varepsilon^{1/3}$$

as claimed. □

G FURTHER EXPERIMENTAL RESULTS

Here, we provide plots akin to those in Figure 3 for other games and solution concepts. For a description of the solution concepts used in these plots, see Zhang & Sandholm (2022). We experiment on four standard benchmark games, which are the same ones used in by Zhang et al. (2023).

- **Kuhn poker**. We use the three-player version of this standard benchmark introduced by Kuhn (1950).

- **Sheriff**. This game, introduced as a benchmark for correlation in extensive-form games by Farina et al. (2019), is a simplified version of the *Sheriff of Nottingham* board game. A *Smuggler*—who is trying to smuggle illegal items in their cargo—and the *Sheriff*—whose goal is stopping the Smuggler. Further details on the game can be found in Farina et al. (2019).

  The Smuggler first chooses a number $n \in \{0, 1\}$ of illegal items to load on the cargo. Then, the Sheriff decides whether to inspect the cargo. If they choose to inspect, and find illegal goods, the Smuggler has to pay n to the Sheriff. Otherwise, the Sheriff compensates the Smuggler with a reward of 1. If the Sheriff decides not to inspect the cargo, the Sheriff's utility is 0, and the Smuggler's utility is $5n$. After the Smuggler has loaded the cargo, and before the Sheriff decides whether to inspect, the Smuggler can attempt to bribe the Sheriff. To do so, they engage in 2 rounds of bargaining and, for each round i, the Smuggler proposes a bribe $b_i \in \{0, 1, 2\}$, and the Sheriff accepts or declines it. Only the proposal and response from the final round are executed. If the Sheriff accepts a bribe b_2 then they get b_2, while the Smuggler's utility is $5n - b_2$.

- **Battleship**. This game, introduced as a benchmark for correlation in extensive-form games by Farina et al. (2019), is a general-sum version of the classic game Battleship, where two players take turns placing ships of varying sizes and values on two separate grids of size 2×2, and then take turns firing at their opponent. Ships which have been hit at all their tiles are considered destroyed. The game ends when one player loses all their ships, or after each player has fired 2 shots. Each player's payoff is determined by the sum of the value of the opponent's destroyed ships minus two times the number of their own lost ships.

- **Ridesharing**. A benchmark introduced in Zhang et al. (2022). Two drivers compete to serve requests on a road network, an undirected graph $G^{\text{U}} = (V^{\text{U}}, E^{\text{U}})$ depicted in Figure 5 with unit edge cost. Each vertex $v \in V^{\text{U}}$ corresponds to a ride request to be served. Each request has a reward in $\mathbb{R}_{\ge 0}$, which is shown in set notation at vertices in the graph. The first driver arriving at node $v \in V^{\text{U}}$ serves the ride and receives the associated reward. The game terminates when all nodes have been cleared, or after $T = 2$. If the two drivers arrive simultaneously on the same vertex they both get reward 0. Final driver utility is computed as the sum of the rewards obtained from the beginning until the end of the game.

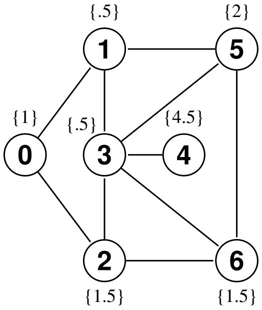


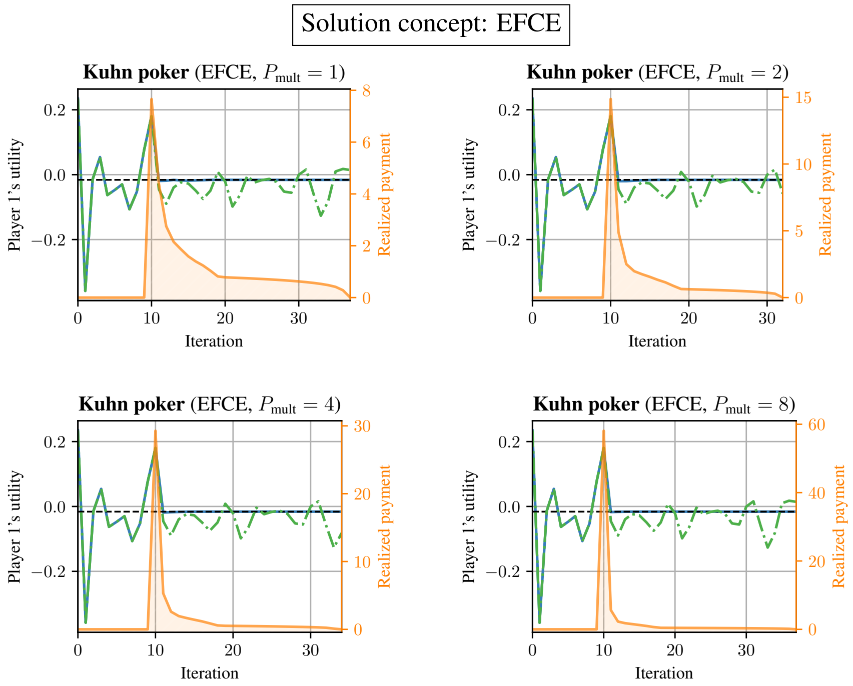
Figure 5: Map used in the ridesharing game. Rewards are in curly braces.

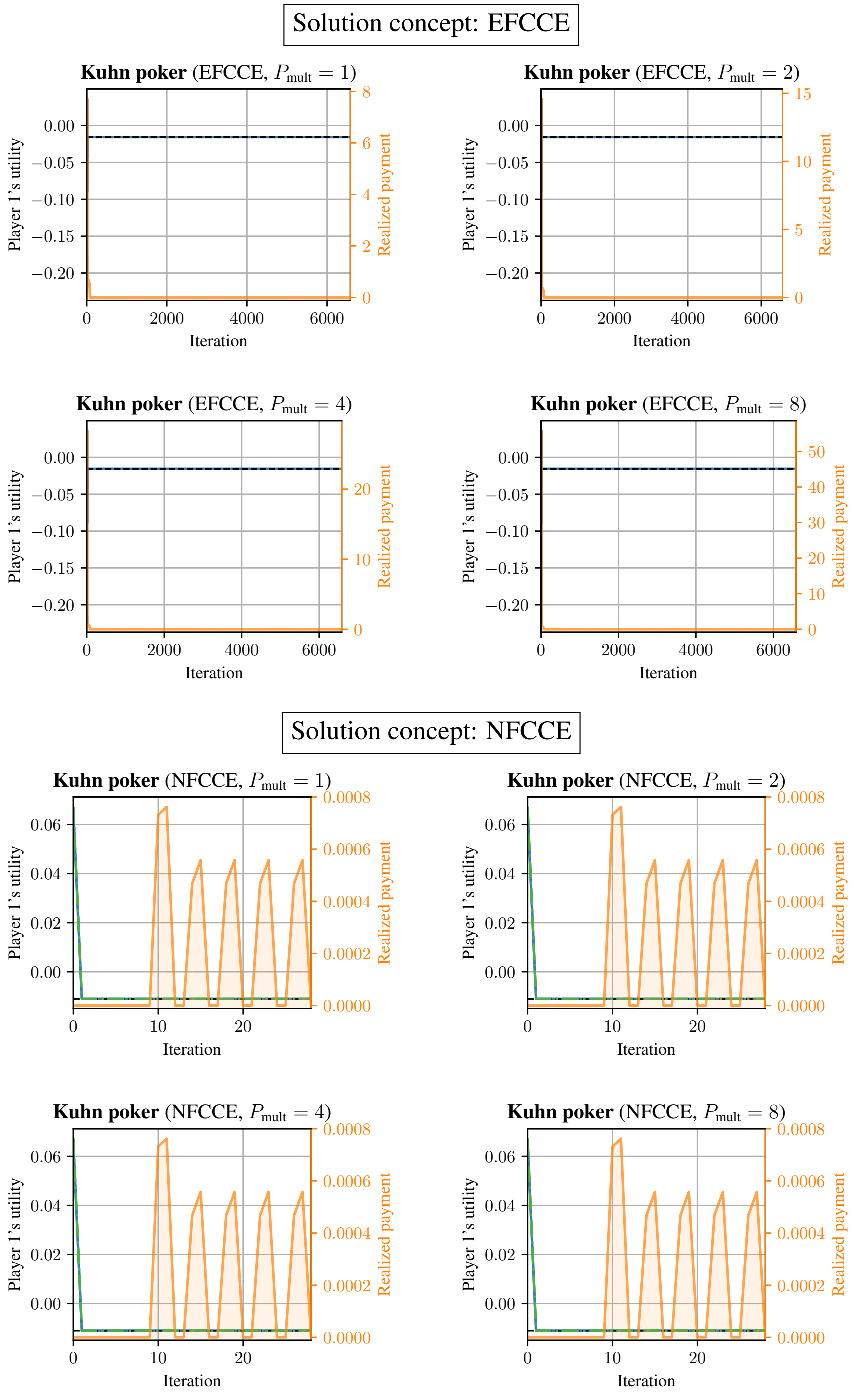
For the following results, we use a burn-in of 10 iterates (that is, no payments are issued in the first 10 iterations; steering only begins after that).

For each game, we consider the problem of steering the learners towards an optimal instance of each of the solution concepts. The objective function used to define optimality is set to be social welfare for general-sum games, and the utility of Player 1 for the three-player zero-sum game (Kuhn poker). For each combination of game and equilibrium concept, we show four plots. Each corresponding to a different value of the payment multiplyer $P_{\text{mult}} \in \{1, 2, 4, 8\}$. The payment multiplier controls the value of P, which is set to $P \coloneqq P_{\text{mult}} \times$ the reward range of the game.

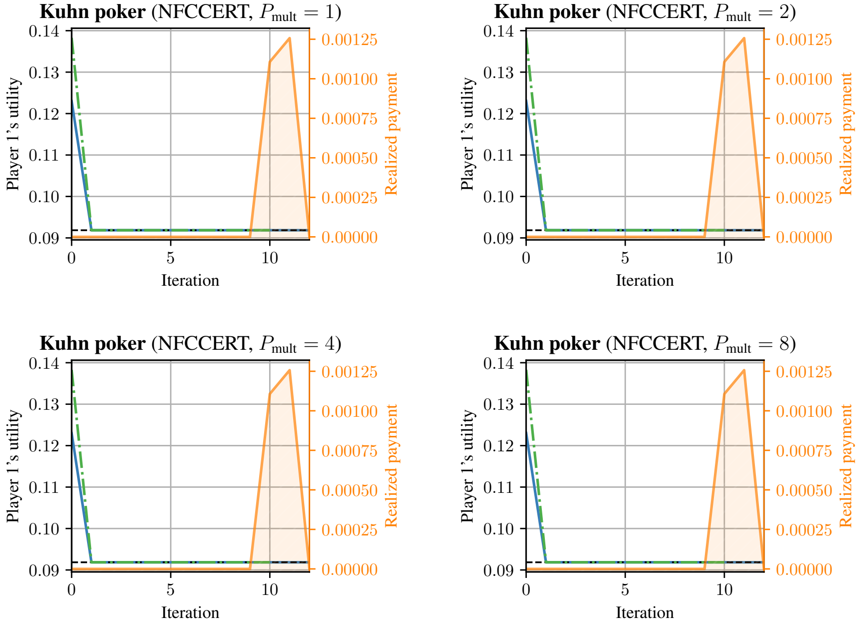
We observe that in all games and equilibrium concepts, our algorithm is able to steer the learners towards the optimal social objective, as predicted by our theory. As P_{mult} grows, we observe that the convergence speed increases, at the cost of a higher payment magnitude.

G.1 Game: Kuhn Poker

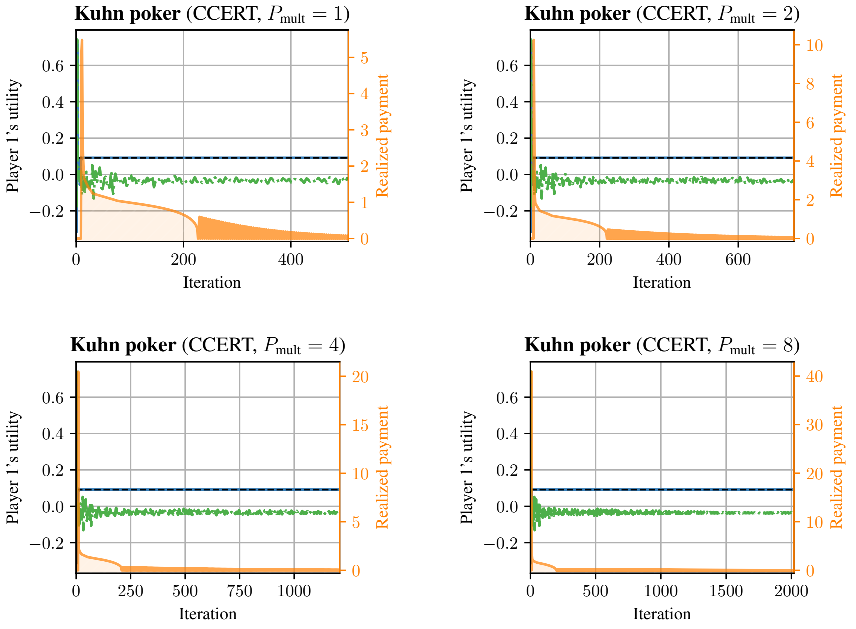



Solution concept: EFCCE
Kuhn poker (EFCCE, $P_{\text{mult}} = 1$)
Kuhn poker (EFCCE, $P_{\text{mult}} = 2$)
Kuhn poker (EFCCE, $P_{\text{mult}} = 4$)
Kuhn poker (EFCCE, $P_{\text{mult}} = 8$)
Player 1's utility
Realized payment
Iteration
0.00
−0.05
−0.10
−0.15
−0.20
0
2000
4000
6000
0
2
4
6
8
5
10
15
20
30
40
50
Solution concept: NFCCE
Kuhn poker (NFCCE, $P_{\text{mult}} = 1$)
Kuhn poker (NFCCE, $P_{\text{mult}} = 2$)
Kuhn poker (NFCCE, $P_{\text{mult}} = 4$)
Kuhn poker (NFCCE, $P_{\text{mult}} = 8$)
0.06
0.04
0.02
0.00
0.0008
0.0006
0.0004
0.0002
0.0000
0
10
20

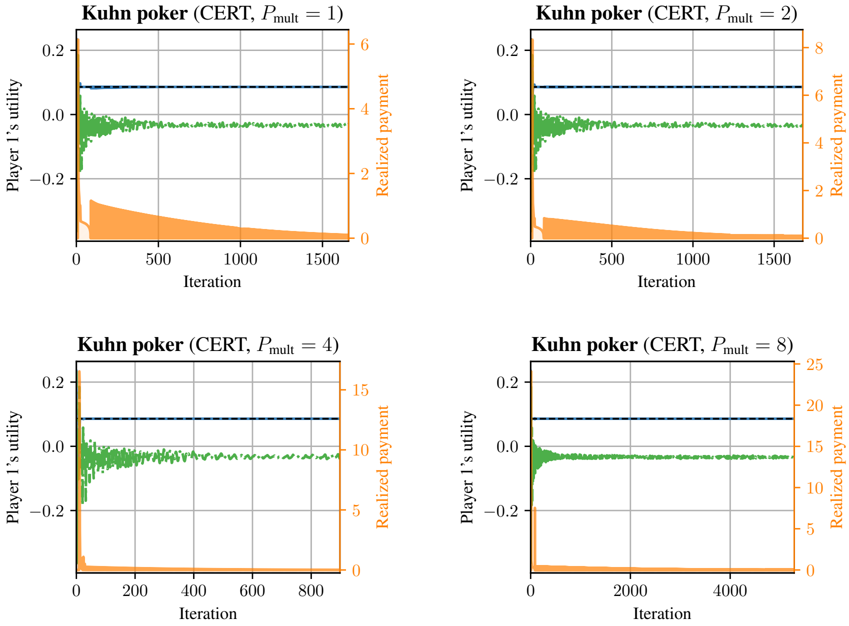
Solution concept: NFCCERT



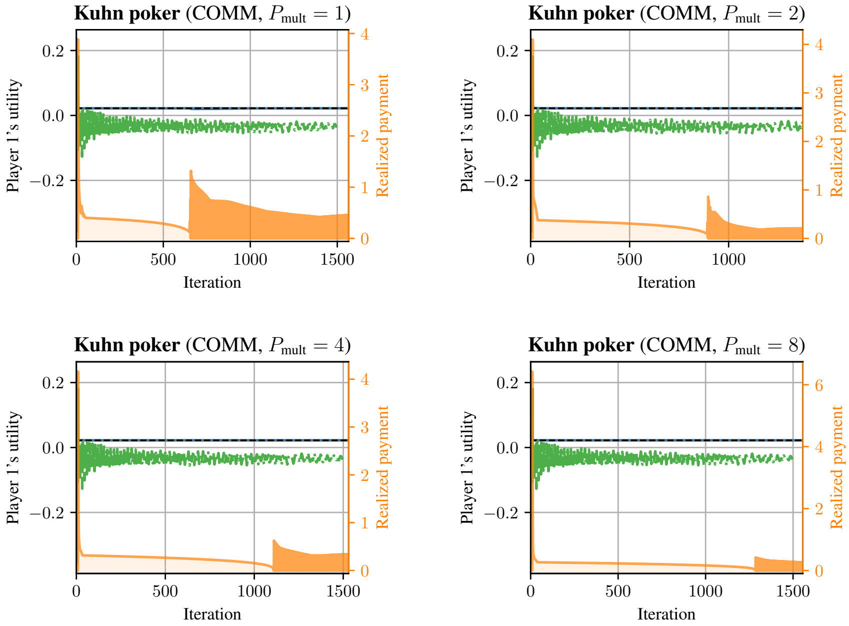
Solution concept: CCERT



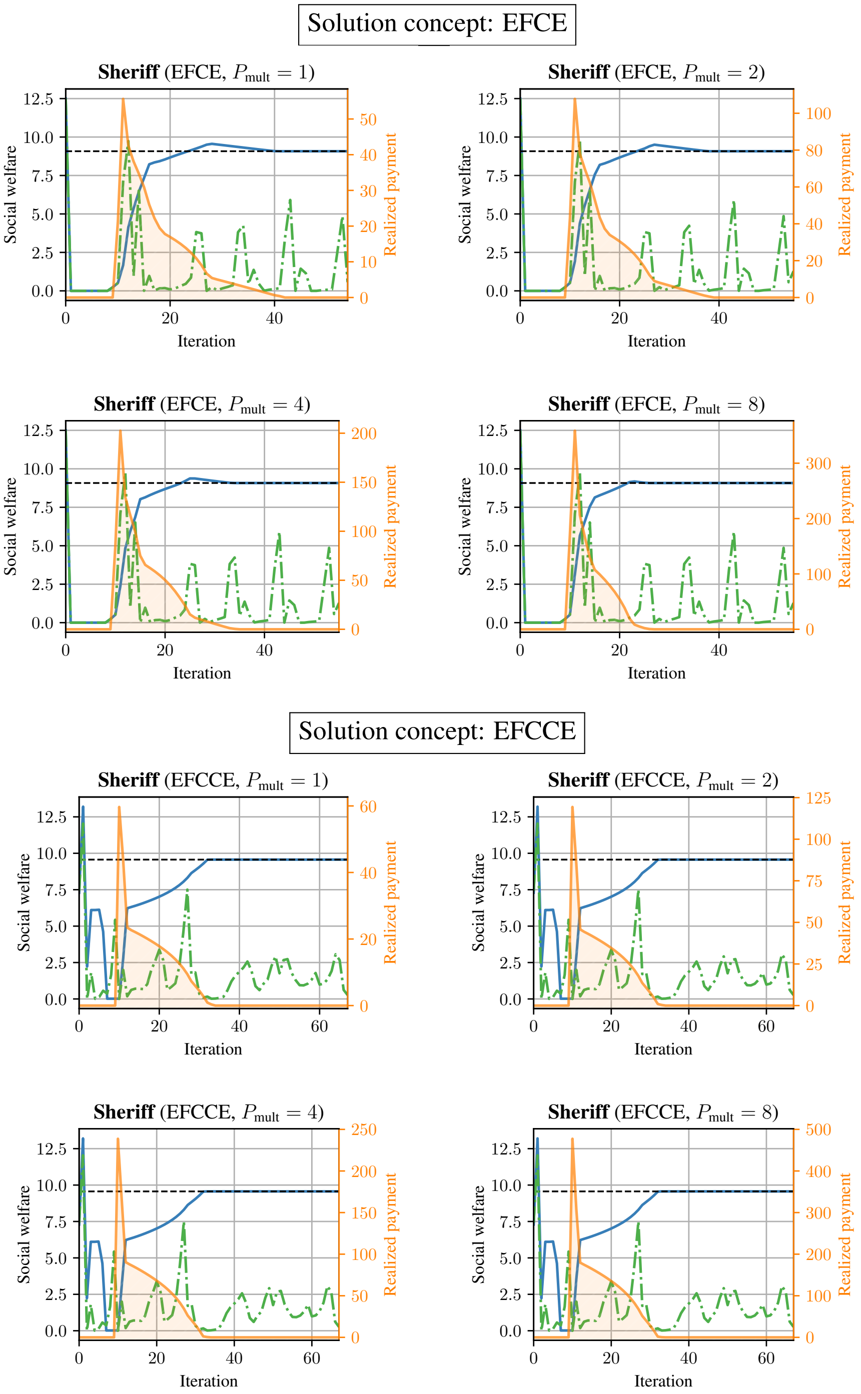
Solution concept: CERT

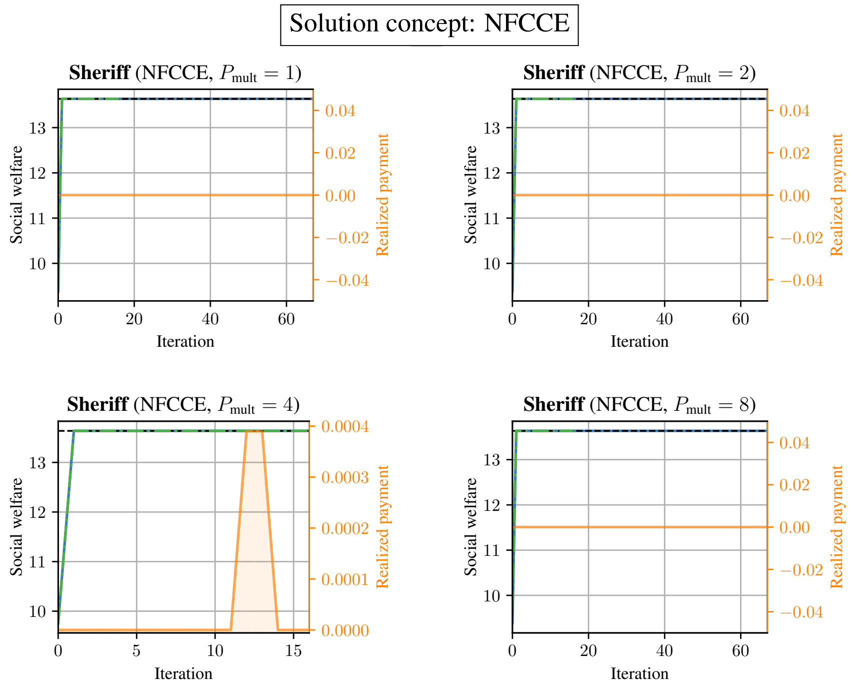


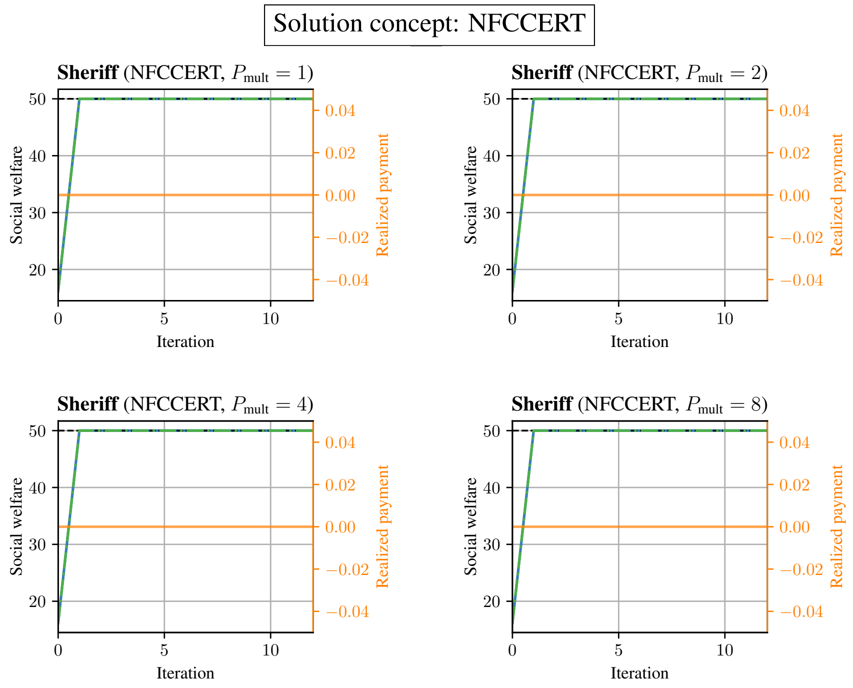
Solution concept: COMM

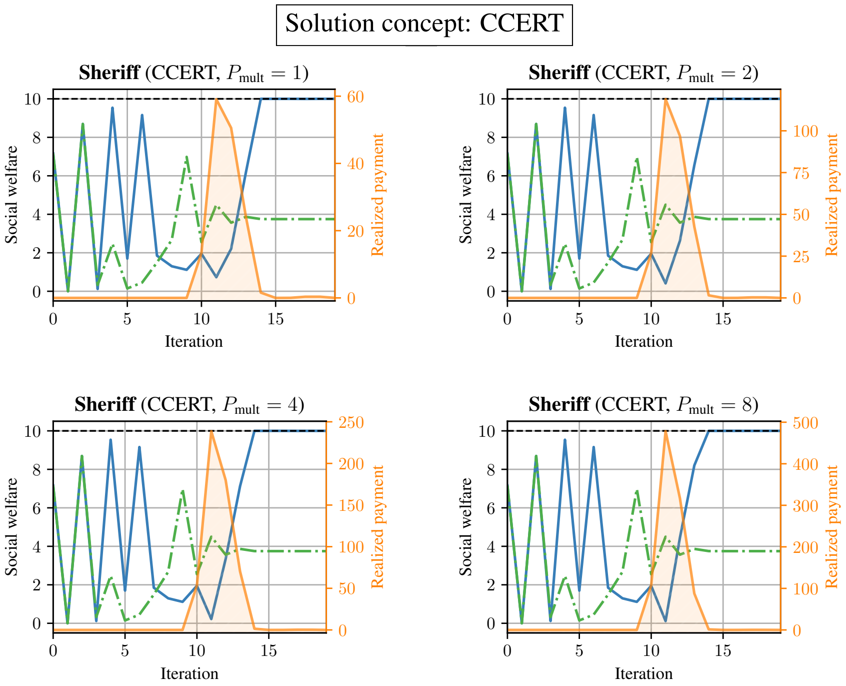


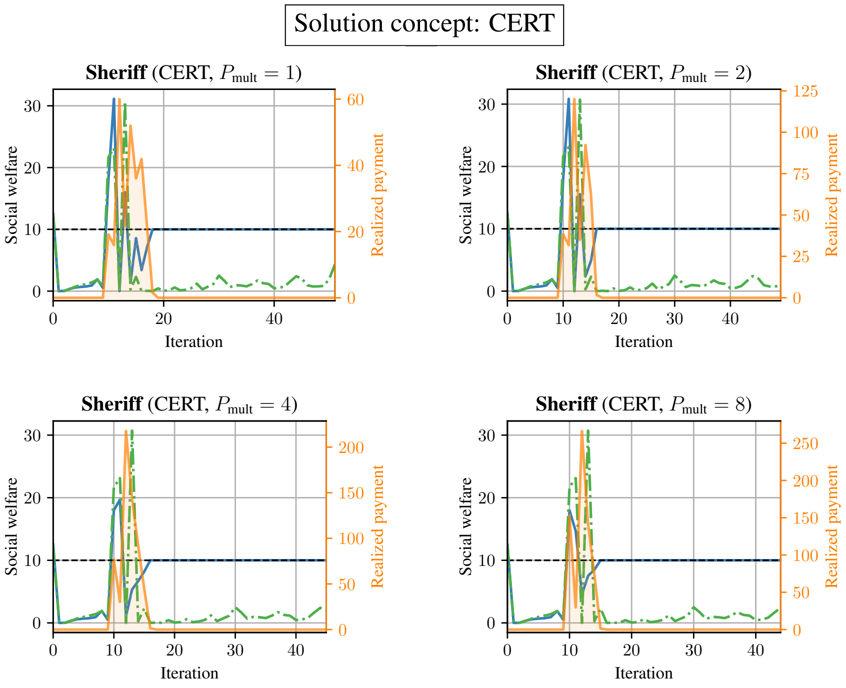
G.2 GAME: SHERIFF



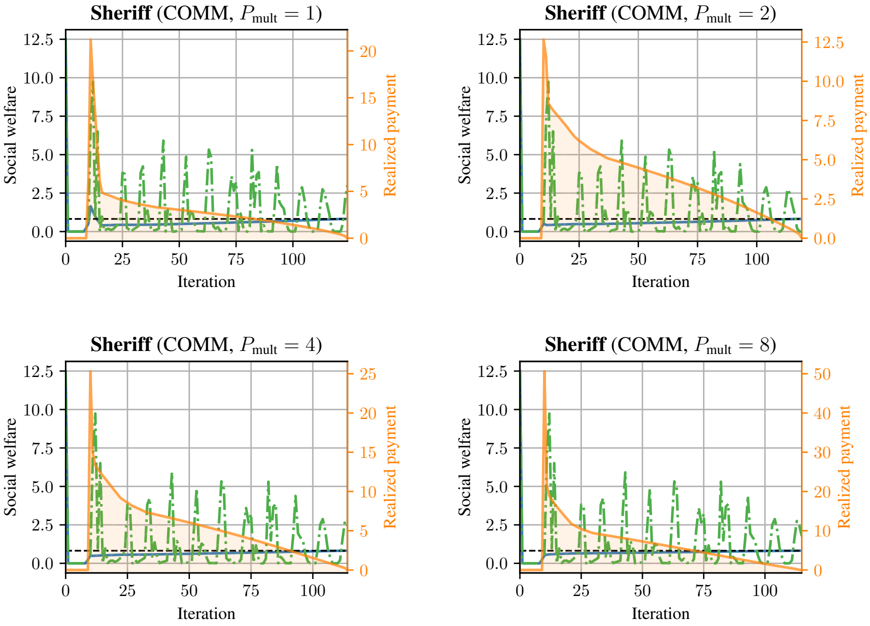

Solution concept: NFCCE
Sheriff (NFCCE, $P_{\text{mult}} = 1$)
Sheriff (NFCCE, $P_{\text{mult}} = 2$)
Sheriff (NFCCE, $P_{\text{mult}} = 4$)
Sheriff (NFCCE, $P_{\text{mult}} = 8$)
Social welfare
10
11
12
13
Realized payment
−0.04
−0.02
0.00
0.02
0.04
0.0000
0.0001
0.0002
0.0003
0.0004
Iteration
0
20
40
60
5
10
15


Solution concept: NFCCERT
Sheriff (NFCCERT, $P_{\text{mult}} = 1$)
Sheriff (NFCCERT, $P_{\text{mult}} = 2$)
Sheriff (NFCCERT, $P_{\text{mult}} = 4$)
Sheriff (NFCCERT, $P_{\text{mult}} = 8$)
Social welfare
20
30
40
50
Realized payment
−0.04
−0.02
0.00
0.02
0.04
Iteration
0
5
10


Solution concept: CCERT
Sheriff (CCERT, $P_{\text{mult}} = 1$)
Sheriff (CCERT, $P_{\text{mult}} = 2$)
Sheriff (CCERT, $P_{\text{mult}} = 4$)
Sheriff (CCERT, $P_{\text{mult}} = 8$)
Social welfare
Realized payment
Iteration
0
2
4
6
8
10
5
15
20
25
40
50
60
75
100
150
200
250
300
400
500


Solution concept: CERT
Sheriff (CERT, $P_{\text{mult}} = 1$)
Sheriff (CERT, $P_{\text{mult}} = 2$)
Sheriff (CERT, $P_{\text{mult}} = 4$)
Sheriff (CERT, $P_{\text{mult}} = 8$)
Social welfare
Realized payment
Iteration
0
10
20
30
40
25
50
60
75
100
125
150
200
250

Solution concept: COMM



Sheriff (COMM, $P_{\text{mult}} = 1$)

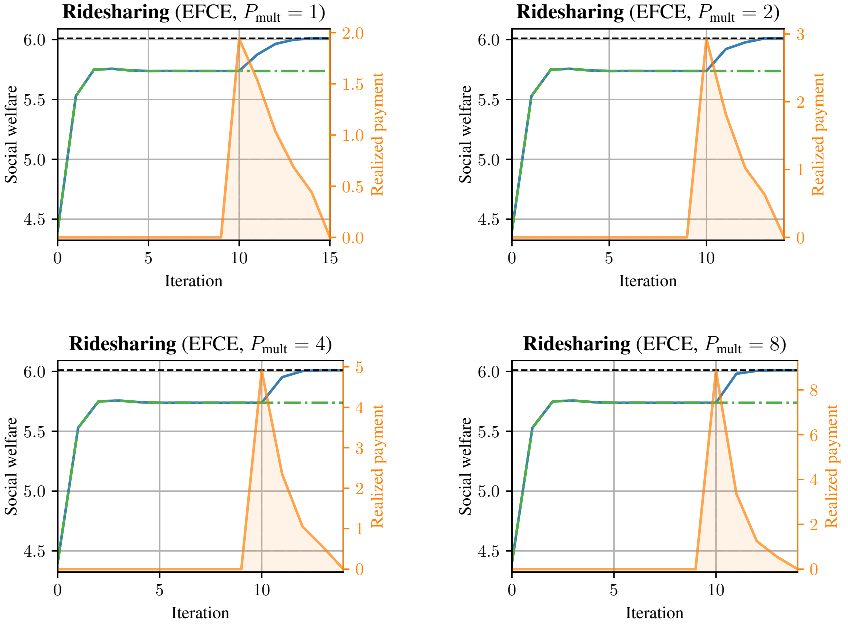
Sheriff (COMM, $P_{\text{mult}} = 2$)

Sheriff (COMM, $P_{\text{mult}} = 4$)

Sheriff (COMM, $P_{\text{mult}} = 8$)

G.3 Game: Ridesharing

Solution concept: EFCE

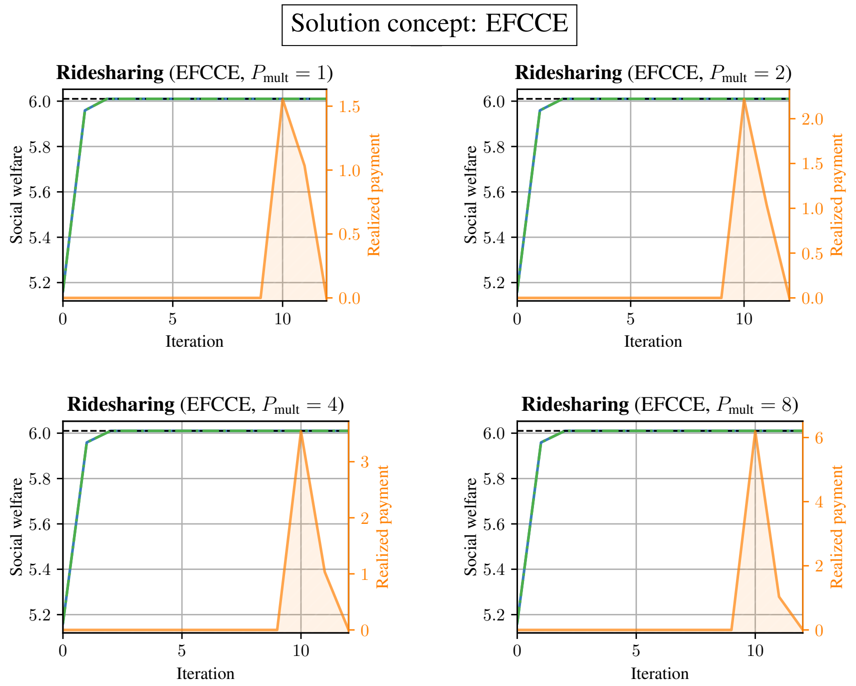


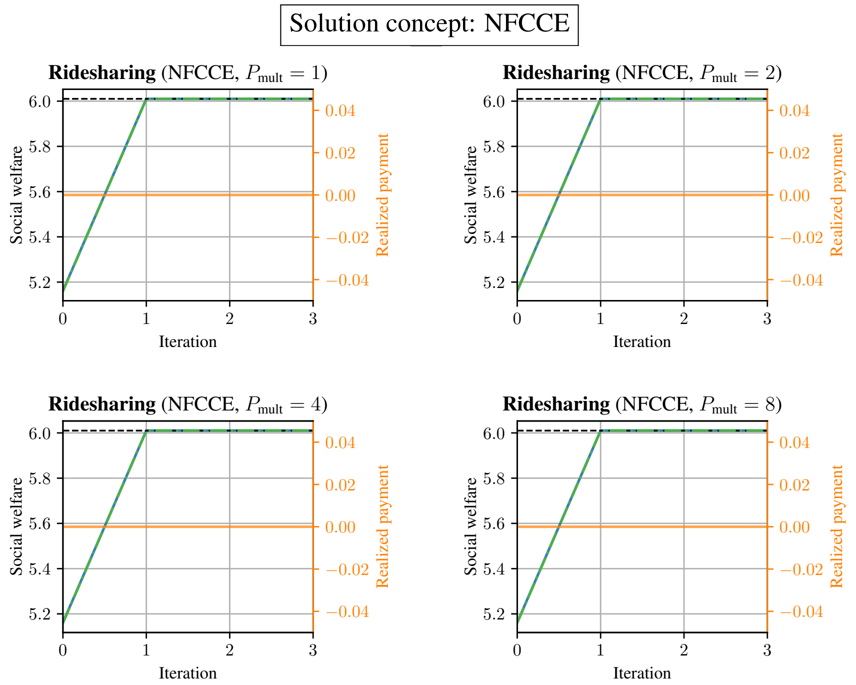
Ridesharing (EFCE, $P_{\text{mult}} = 1$)

Ridesharing (EFCE, $P_{\text{mult}} = 2$)

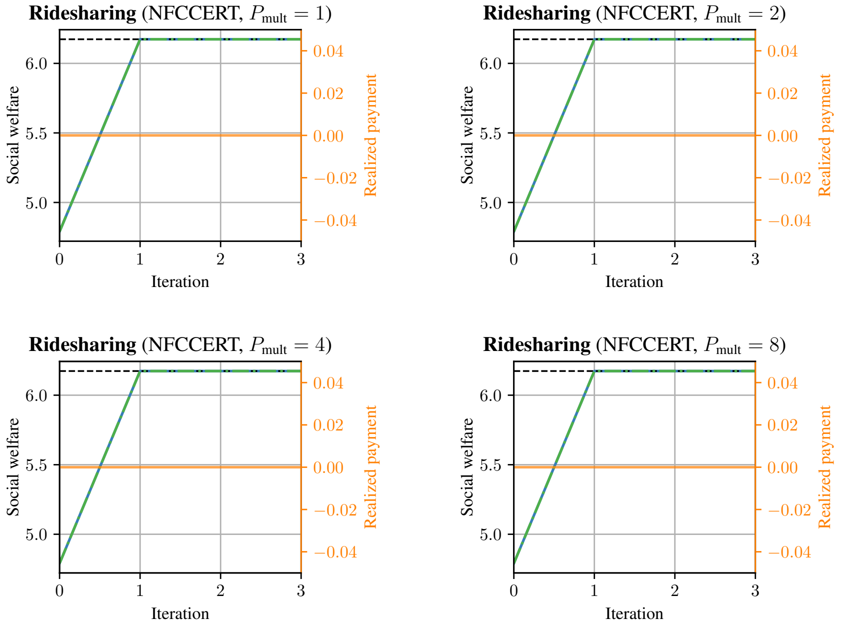
Ridesharing (EFCE, $P_{\text{mult}} = 4$)

Ridesharing (EFCE, $P_{\text{mult}} = 8$)

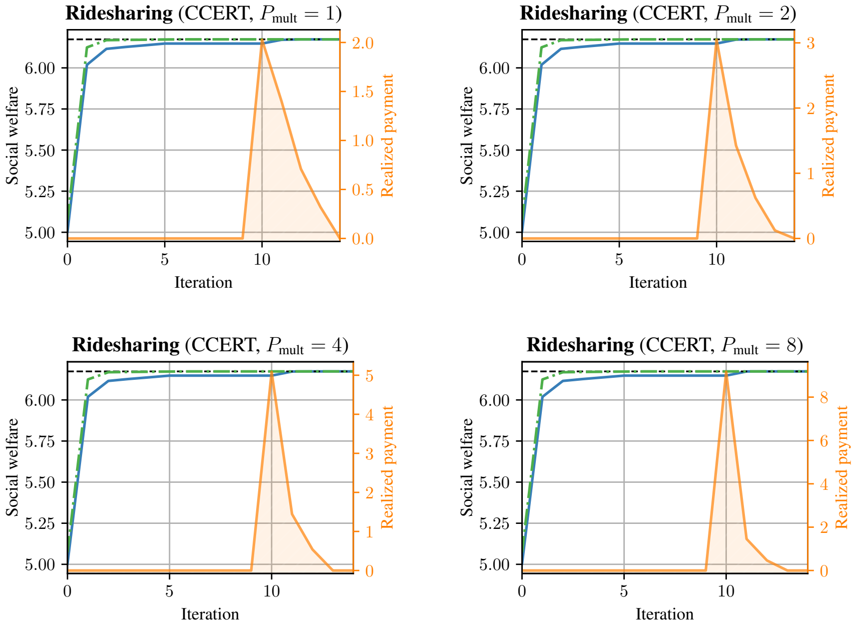

Solution concept: EFCCE
Ridesharing (EFCCE, $P_{\text{mult}} = 1$)
Ridesharing (EFCCE, $P_{\text{mult}} = 2$)
Ridesharing (EFCCE, $P_{\text{mult}} = 4$)
Ridesharing (EFCCE, $P_{\text{mult}} = 8$)
Social welfare
Realized payment
Iteration
5.2
5.4
5.6
5.8
6.0
0
5
10
0.0
0.5
1.0
1.5
2.0
1
2
3
4
6


Solution concept: NFCCE
Ridesharing (NFCCE, $P_{\text{mult}} = 1$)
Ridesharing (NFCCE, $P_{\text{mult}} = 2$)
Ridesharing (NFCCE, $P_{\text{mult}} = 4$)
Ridesharing (NFCCE, $P_{\text{mult}} = 8$)
Social welfare
Realized payment
Iteration
5.2
5.4
5.6
5.8
6.0
0
1
2
3
0.04
0.02
0.00
−0.02
−0.04

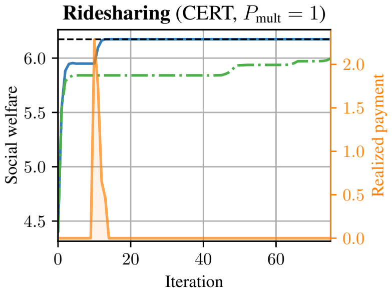
Solution concept: NFCCERT



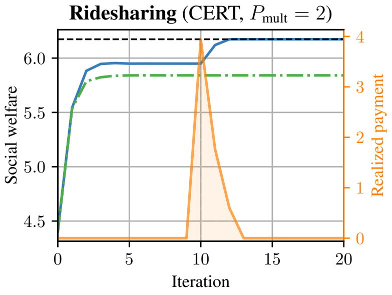
Solution concept: CCERT



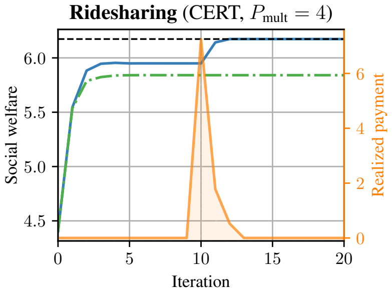
Solution concept: CERT



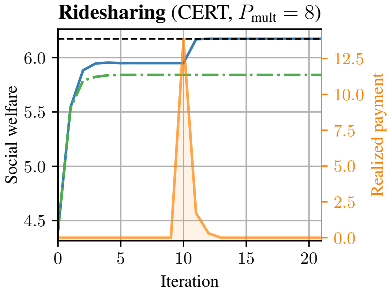
Ridesharing (CERT, $P_{\text{mult}} = 1$)



Ridesharing (CERT, $P_{\text{mult}} = 2$)

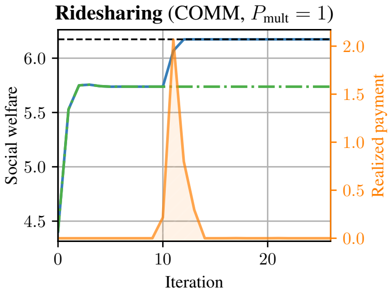


Ridesharing (CERT, $P_{\text{mult}} = 4$)

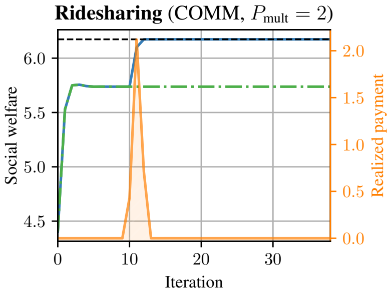


Ridesharing (CERT, $P_{\text{mult}} = 8$)

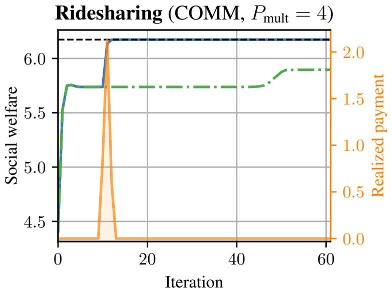
Solution concept: COMM



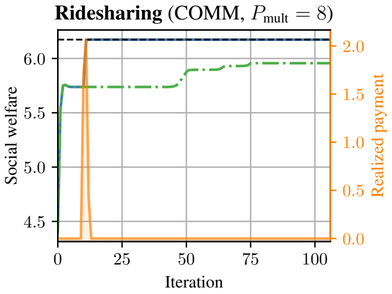
Ridesharing (COMM, $P_{\text{mult}} = 1$)



Ridesharing (COMM, $P_{\text{mult}} = 2$)

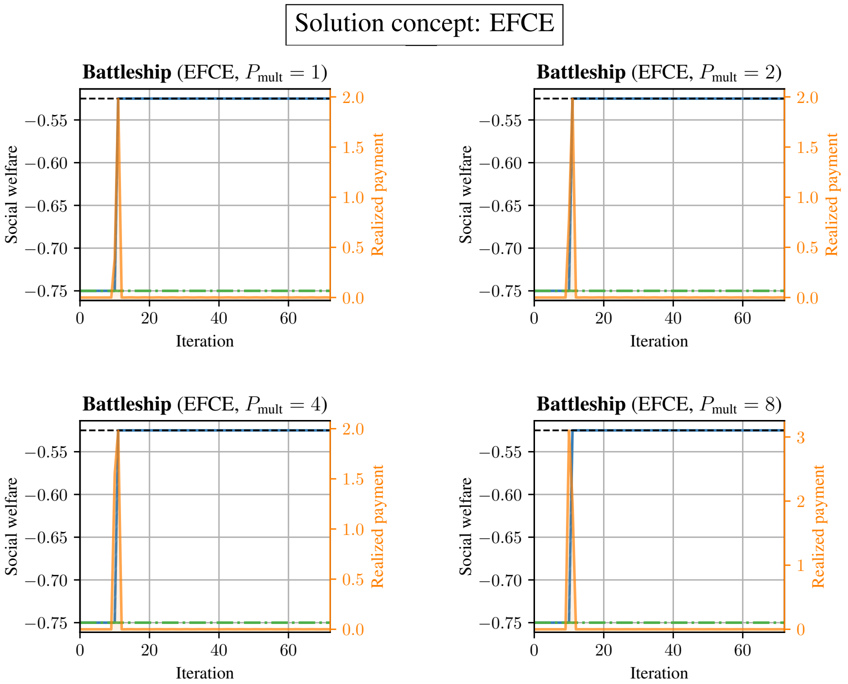


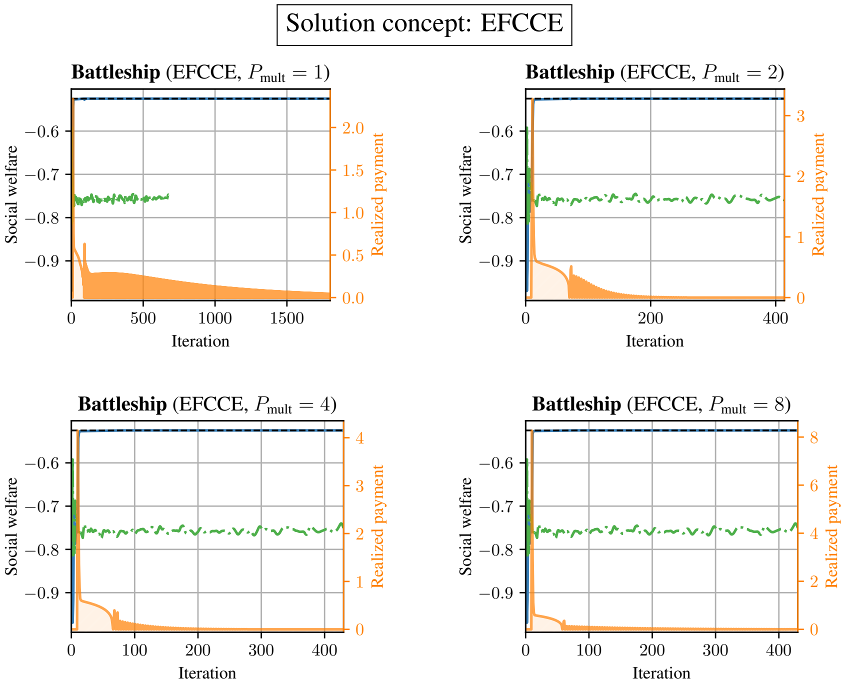
Ridesharing (COMM, $P_{\text{mult}} = 4$)

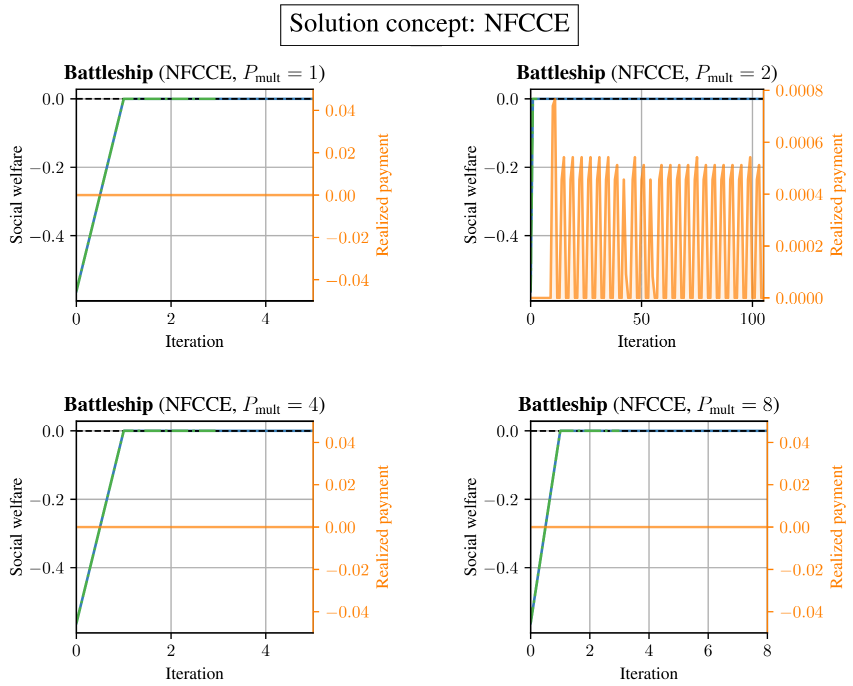


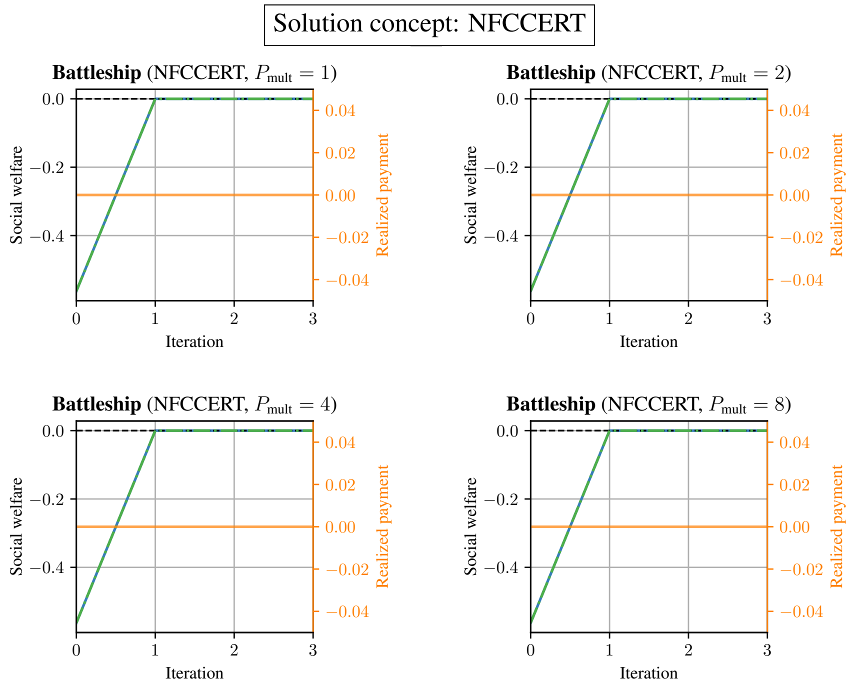
Ridesharing (COMM, $P_{\text{mult}} = 8$)

G.4 GAME: BATTLESHIP

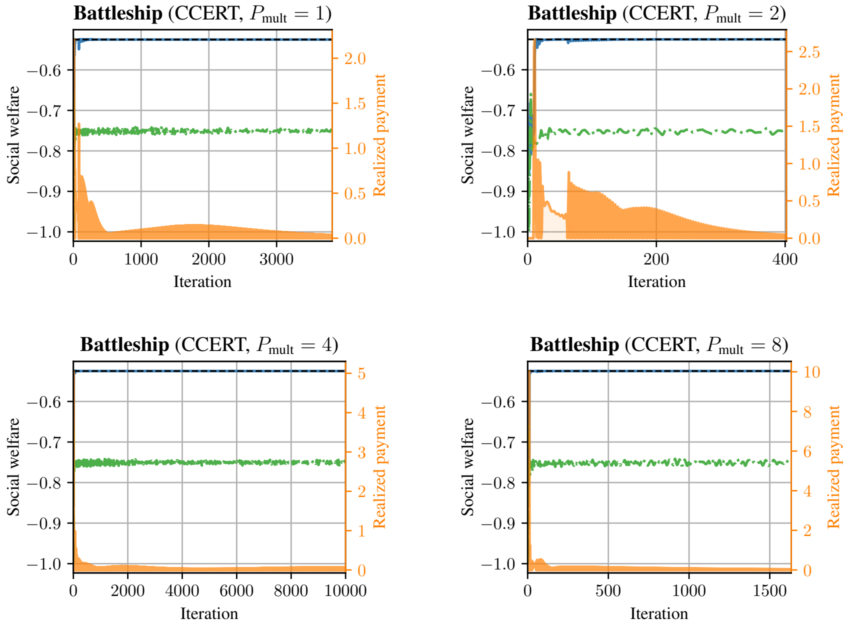




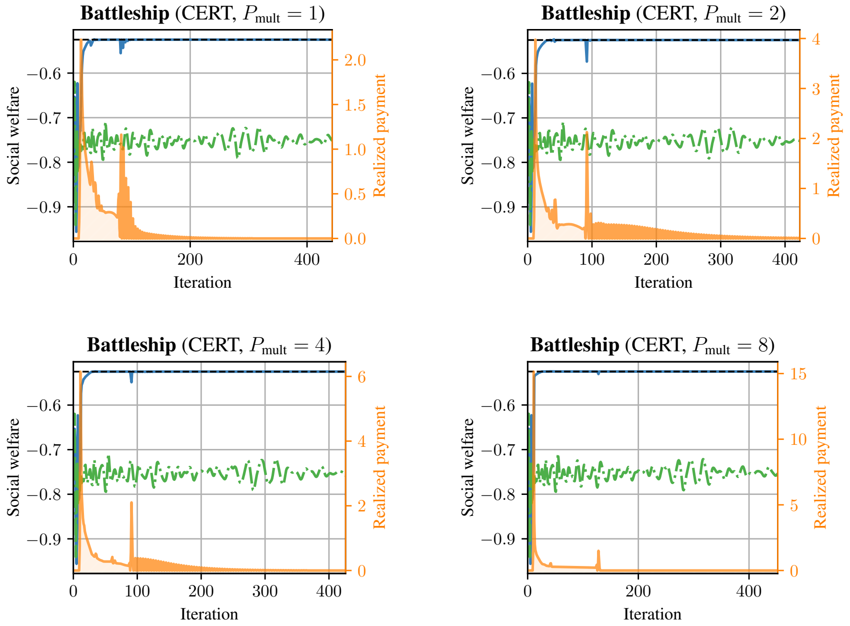

Solution concept: NFCCE
Battleship (NFCCE, $P_{\text{mult}} = 1$)
Battleship (NFCCE, $P_{\text{mult}} = 2$)
Battleship (NFCCE, $P_{\text{mult}} = 4$)
Battleship (NFCCE, $P_{\text{mult}} = 8$)
Social welfare
0.0
−0.2
−0.4
Realized payment
0.04
0.02
0.00
−0.02
−0.04
0.0008
0.0006
0.0004
0.0002
0.0000
Iteration
0
2
4
6
8
50
100


Solution concept: NFCCERT
Battleship (NFCCERT, $P_{\text{mult}} = 1$)
Battleship (NFCCERT, $P_{\text{mult}} = 2$)
Battleship (NFCCERT, $P_{\text{mult}} = 4$)
Battleship (NFCCERT, $P_{\text{mult}} = 8$)
Social welfare
0.0
−0.2
−0.4
Realized payment
0.04
0.02
0.00
−0.02
−0.04
Iteration
0
1
2
3

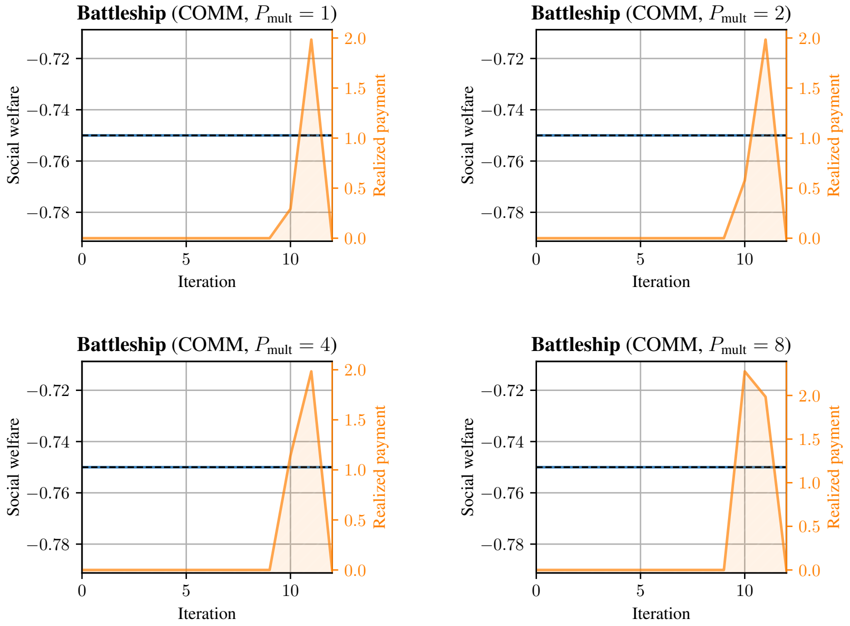
Solution concept: CCERT



Solution concept: CERT



Solution concept: COMM



Battleship (COMM, $P_{\text{mult}} = 1$)

Battleship (COMM, $P_{\text{mult}} = 2$)

Battleship (COMM, $P_{\text{mult}} = 4$)

Battleship (COMM, $P_{\text{mult}} = 8$)