PAC Learning and Stabilizing Hedonic Games: Towards a Unifying Approach.

Simone Fioravanti[1], Michele Flammini[1,4], Bojana Kodric[2,1], Giovanna Varricchio[3]

[1] Gran Sasso Science Institute (GSSI), L'Aquila, Italy
[2] Ca' Foscari University of Venice, Venice, Italy
[3] Goethe-Universität, Frankfurt am Main, Germany
[4] University of Calabria, Rende, Italy
simone.fioravanti@gssi.it, michele.flammini@gssi.it, bojana.kodric@unive.it, varricchio@em.uni-frankfurt.de

Abstract

We study PAC learnability and PAC stabilizability of Hedonic Games (HGs), i.e., efficiently inferring preferences or core-stable partitions from samples. We first expand the known learnability/stabilizability landscape for some of the most prominent HGs classes, providing results for Friends and Enemies Games, Bottom Responsive, and Anonymous HGs. Then, having a broader view in mind, we attempt to shed light on the structural properties leading to learnability/stabilizability, or lack thereof, for specific HGs classes. Along this path, we focus on the fully expressive Hedonic Coalition Nets representation of HGs. We identify two sets of conditions that lead to efficient learnability, and which encompass all of the known positive learnability results. On the side of stability, we reveal that, while the freedom of choosing an ad hoc adversarial distribution is the most obvious hurdle to achieving PAC stability, it is not the only one. First, we show a distribution independent necessary condition for PAC stability. Then, we focus on $\mathcal{W}$-games, where players have individual preferences over other players and evaluate coalitions based on the least preferred member. We prove that these games are PAC stabilizable under the class of bounded distributions, which assign positive probability mass to all coalitions. Finally, we discuss why such a result is not easily extendable to other HGs classes even in this promising scenario. Namely, we establish a purely computational property necessary for achieving PAC stability.

Introduction

Hedonic Games (HGs) (Dreze and Greenberg 1980) are a formal model for describing selfish individuals gathering together in order to form coalitions. Both HGs and general coalition formation games attracted considerable research attention in the last years due to their applicability to multi-agent environments. Solution concepts for HGs are usually in the form of agent partitions with some suitable properties. The one we consider in this paper is core stability. A partition is said to be *core-stable* (or in the core) if there exists no subset of players that could regroup into a so-called *core-blocking* coalition, which is preferred by all of them.

The usual assumption when considering any solution concept is that the preferences of the agents are fully known, which is arguably unrealistic. Could we instead efficiently

infer the whole game structure, or even directly learn solution concepts, while having only partial knowledge of the preferences? Questions of this kind are naturally captured by the *probably approximately correct (PAC) learning* framework (Valiant 1984), which formalizes the problem of learning a target concept from a limited number of samples from any possible unknown but fixed distribution.

Sliwinski and Zick (2017) were the first to leverage the PAC framework to study the problem of learning HGs preferences and core-stable partitions from samples. In particular, they define PAC stabilizability of a HGs class as the property of being able to, upon seeing a limited number of samples, either report that the core is empty or propose a partition that is unlikely to be core-blocked by further coalitions sampled from the same distribution. In a recent paper, Lev et al. (2021a) apply the notion of PAC stabilizability of HGs in the context of political coalition formation. In particular, they use the publicly available Israeli parliament voting data to fit a Friends Appreciation HG, and compare the actual political parties of the voters to the PAC-stable coalitions resulting from the model. This example shows how learning concepts have the potential to create space for applications of mainly theoretical models, as HGs.

While the work of Sliwinski and Zick (2017) and the ones that followed considered PAC learnability and stabilizability of many specific classes of HGs, the overall picture is still far from being complete. Most prominently, the characterization of the underlying general conditions explaining the existing results is missing. Furthermore, PAC stabilizability seems very hard to achieve and it is natural to wonder whether some restrictions on the PAC stability definition can yield better results. Here, we address these questions, attempting to provide a deeper theoretical understanding of what makes HGs learnable and stabilizable.

Our Contribution

We first extend the knowledge on PAC learnable and PAC stabilizable classes of HGs. We start by focusing on Friends and Enemies Games, examining whether the negative results on stabilizability of Additively Separable HGs transfer to this simple subclass. By exploiting previous results and proposing an algorithm stabilizing Friends and Enemies under Enemies Aversion, we deduce that Friends and Enemies Games belong to the very few lucky HGs classes that can

both be learned and stabilized. Next, we study Bottom Responsive HGs and show that while they are not efficiently learnable, they are stabilizable. Finally, we turn our attention to Anonymous HGs and show that the opposite holds here, i.e., they are efficiently learnable but not stabilizable.

After exploring specific HGs classes, we use the gained insights to follow a more general research direction, devoted to a deeper understanding of the structural properties that make HGs learnable and/or stabilizable.

We first consider the learning problem. Additively Separable, Anonymous, $\mathcal{W}$ and $\mathcal{B}$-games are all known to be learnable, and we investigate why this is the case. To this aim, we consider Hedonic Coalition Nets (HCNs), a general framework for representing HGs that is universally expressive, i.e., it can represent any HGs class. We identify two sets of conditions on the HCNs representation that imply efficient learnability, and as special cases explain the learnability of all of the aforementioned HGs classes.

We then turn our attention to stability. Achieving PAC stability does not seem possible for most HGs classes, and we try to find general reasons causing this fact. First, we show a simple necessary condition for PAC stability, abstracting the proof pattern of all the known negative results for specific HGs classes. Then, we consider the problem of PAC stability with bounded probability distributions and prove that under this restriction it is possible to PAC stabilize $\mathcal{W}$-games, which is known not to be possible in general. Finally, we discuss why the same result cannot be easily extended to other HGs. In particular, we determine a general purely computational property necessary for achieving PAC stability.

Due to space limitations, some proofs and details are omitted. We refer the interested reader to the full version of the paper.

Related Work

Many works have dealt with learning game-theoretic solution concepts from data. Sliwinski and Zick (2017) first introduced the PAC learning framework into the study of HGs. Their work was extended by Igarashi, Sliwinski, and Zick (2019) to tackle HGs with underlying players' interaction networks. Moreover, Jha and Zick (2020) laid further foundations for learning game-theoretic solution concepts from samples. More recently, Trivedi and Hemachandra (2021) studied the problem of learning HGs with noisy preferences.

Other works have considered learning cooperative games (Balcan, Procaccia, and Zick 2015), markets (Lev et al. 2021b), auctions (Balcan, Sandholm, and Vitercik 2018) but also, more generally, combinatorial functions (Balcan, Vitercik, and White 2016; Balcan 2015).

There is a vast body of literature on HGs. For a thorough introduction to the main concepts and results, we refer to Aziz and Savani (2016), where both all the HGs classes studied in this paper and also HCNs are discussed.

Preliminaries

Let N be a set of n *players*. We call any non-empty subset $S \subseteq N$ a *coalition* and denote by $\mathcal{N}_i$ the set of all coalitions which contain a given player $i \in N$. We call any coalition of size one a *singleton*. We denote by $\succsim_i$ any binary *preference relation* of player i over the coalitions in $\mathcal{N}_i$, which is reflexive, transitive, and complete. A *Hedonic Game* (HG) is then a pair $H = (N, \succsim)$, where $\succsim = (\succsim_i, \ldots, \succsim_n)$ is a *preference profile*, i.e., the collection of all players' preferences. Throughout this work we will assume that players' preferences are expressed as real numbers by means of *valuation* functions v_i. In other words, given $S, T \in \mathcal{N}_i$: $v_i(S) \geq v_i(T)$ if and only if $S \succsim_i T$. We will denote by $\vec{v} = (v_1, \ldots, v_n)$ the collections of players' valuations and assume that $v_i(S) = \varnothing$ for $S \notin \mathcal{N}_i$. Let H be a HG and π a *coalition structure*, i.e., a partition of players into coalitions. A set S is said to *core-block* π if $v_i(S) > v_i(\pi(i))$ for each $i \in S$, where $\pi(i)$ denotes the coalition containing i in π. A coalition structure π is *core-stable* if there does not exist a core-blocking coalition $S \subseteq N$. Among the many possible solution concepts, the one we will consider in this paper is core stability, as it is the most prominent one in the PAC stability model.

Defining Classes of Hedonic Games

In this subsection, we provide the definitions of some HGs classes already considered from the perspective of PAC learning by (Sliwinski and Zick 2017), that will be frequently mentioned in the sequel. In all of these classes, for a player $i \in N$ and a coalition $S \in \mathcal{N}_i$, the valuation $v_i(S)$ is completely determined by the values $v_i(j)$ for $j \in S \setminus \{i\}$. More precisely, the valuation of i for S is equal to:

1. *Additively Separable*: the sum of the values of its members, i.e., $v_i(S) = \sum_{j \in S \setminus \{i\}} v_i(j)$;

2. *Fractional*: the sum of the values of its members, but normalized by the size of the coalition, i.e., $v_i(S) = \sum_{j \in S \setminus \{i\}} v_i(j) / |S|$;

3. $\mathcal{W}$-*games*: the value of the worst player in the coalition;

4. $\mathcal{B}$-*games*: the value of the best player in the coalition, but coalitions of smaller size are preferred.

PAC Learning

The PAC learning model, originally introduced by Valiant (1984), mathematically formalizes the process of learning a *target concept* v belonging to a *hypothesis class* $\mathcal{H}$, by using a sample of labeled examples as input. There are many variants, which adapt to different learning paradigms. In the following, we will formally present only the one that we will use in this work. Our aim is to learn an unknown valuation function $v : 2^N \to \mathbb{R}$ within a class $\mathcal{H}$, given as input $\mathcal{S} = \{(S_1, v(S_1)), \ldots, (S_m, v(S_m))\}$, i.e., a collection of coalition/valuations pairs. The distribution $\mathcal{D}$, according to which the i.i.d. input coalitions are sampled, is unknown, while the class $\mathcal{H}$ is determined by the HG instance one considers, e.g., if one is studying Additively Separable HGs, $\mathcal{H}$ will be the class of additively separable functions over (the other) $n - 1$ players. Starting from a sample $\mathcal{S}$, learning is the process of producing a hypothesis $v^* \in \mathcal{H}$ which is as close as possible to the real v. Formally, a hypothesis $v^* \in \mathcal{H}$ is ε-*approximately correct* w.r.t. a distribution $\mathcal{D}$ over 2^N and

a function $v \in \mathcal{H}$, if the following holds:

$$\Pr_{S \sim \mathcal{D}} \left[v^*(S) \neq v(S) \right] < \varepsilon .$$

Given $\varepsilon, \delta > 0$, class $\mathcal{H}$ is (ε, δ) *probably approximately correctly (PAC) learnable* if there exists an algorithm $\mathcal{A}$ that, for every distribution $\mathcal{D}$ over 2^N, and any $v \in \mathcal{H}$, given a sample drawn from $\mathcal{D}$, is able to produce a hypothesis v^* which is ε-approximately correct with probability at least $1 - \delta$. A class $\mathcal{H}$ is said to be PAC learnable if it is (ε, δ) PAC learnable for all $\varepsilon, \delta > 0$. Furthermore, if the sample size m and the running time of $\mathcal{A}$ are polynomial in $\frac{1}{\varepsilon}, \log \frac{1}{\delta}$ and n, $\mathcal{H}$ is said to be *efficiently PAC learnable*.

The inherent complexity of efficiently PAC learning a concept class of real functions $\mathcal{H}$ is usually measured by the so-called *pseudo-dimension* (see, e.g., Anthony and Bartlett (2002)), which is the analog of the more renowned *VC-dimension* (Kearns and Vazirani 1994) defined only for classes of binary functions. In order to formally define pseudo-dimension, we first need to introduce the concept of *pseudo-shattering*. Given a collection of coalition/value pairs $\mathcal{S} = \{(S_1, r_1), \ldots, (S_q, r_q)\}$, we say that a class $\mathcal{H}$ can *pseudo-shatter* $\mathcal{S}$ if, for every possible binary labeling $l_1, \ldots, l_q$ of $\mathcal{S}$, there exists a function $f \in \mathcal{H}$ such that $f(S_j) > r_j \iff l_j = 1$. Intuitively, the more $\mathcal{H}$ is expressive, the bigger the sets that it can pseudo-shatter. The pseudo-dimension of $\mathcal{H}$, denoted as $P_{dim}(\mathcal{H})$, is the size of the maximal set $\mathcal{S}$ that can be pseudo-shattered by $\mathcal{H}$.

We conclude this section by reporting the theorem which bridges learning and pseudo-dimension.

Theorem 1 (Anthony and Bartlett 2002). *A hypothesis class $\mathcal{H}$ with $P_{dim}(\mathcal{H})$ polynomial in n is (ε, δ) PAC learnable using m samples, where m is polynomial in $P_{dim}(\mathcal{H}), \frac{1}{\varepsilon}$ and $\log \frac{1}{\delta}$, by any algorithm $\mathcal{A}$ that returns a hypothesis f^* consistent with the sample, i.e., $f^*(S_i) = f(S_i)$ for all i. Furthermore, if $P_{dim}(\mathcal{H})$ is superpolynomial in n, $\mathcal{H}$ is not efficiently PAC learnable.*

PAC Stabilizing Hedonic Games

The concept of *PAC stabilizing* HGs was first introduced in (Sliwinski and Zick 2017). A coalition structure π is said to be ε-*PAC stable* under a distribution $\mathcal{D}$ if $\Pr_{S \sim \mathcal{D}} \left[S \text{ core-blocks } \pi \right] < \varepsilon$. A class of HGs $\mathcal{H}$ is *PAC stabilizable* if there exists an algorithm $\mathcal{A}$ that for any HG in $\mathcal{H}$, any $\varepsilon, \delta > 0$, and any $\mathcal{D}$ over 2^N, given a sample $\mathcal{S} = \{(S_1, \vec{v}(S_1)), \ldots, (S_m, \vec{v}(S_m))\}$ of coalitions drawn according to $\mathcal{D}$, produces an ε-PAC stable coalition structure π under $\mathcal{D}$ with probability at least $1 - \delta$, or reports that the core is empty. If the sample size m and the running time of $\mathcal{A}$ meet the same conditions required for efficient PAC learnability, we say that $\mathcal{H}$ is *efficiently PAC stabilizable*. Intuitively, this concept formalizes the learnability of a solution concept for a HGs class, independently from the learnability of the class itself. We will rely on the following theorem in the next section.

Theorem 2 (Jha and Zick 2020). *A class of HGs $\mathcal{H}$ is efficiently PAC stabilizable iff there exists an algorithm that outputs a partition π consistent with the sample, i.e., no coalition from the sample core-blocks π.*

HGs class	Learnable	Stabilizable
Friends and Enemies		
Friends Appreciation	✓*	✓*
Enemies Aversion	✓*	✓
Bottom Responsive	✗	✓
Anonymous	✓	✗

Table 1: A summary of the learnability and stabilizability landscape discussed below. Entries marked by an asterisk symbol are consequences of previous work.

Learnability and Stabilizability of New Classes of Hedonic Games

In this section we broaden the picture of learnability and stabilizability of different classes of HGs, studying the following HG classes that were not considered by previous work.

Friends and Enemies. Friends and Enemies Games have been traditionally investigated under two types of preference profiles, called *Friends Appreciation* and *Enemies Aversion*, where agents prefer coalitions with a greater number of friends (and smaller number of enemies in case of ties) or with a smaller number of enemies (and greater number of friends in case of ties), respectively.

Bottom Responsive. The *bottom responsiveness* property was first defined by Suzuki and Sung (2010) as bottom *refuseness* and then further considered by Aziz and Brandl (2012), where it was renamed in analogy to a related property called *top responsiveness*. Intuitively, it models pessimistic agents who rank coalitions based on sets of players that they would like to avoid.

Definition 1. *For each player $i \in N$ and $S \in \mathcal{N}_i$, we define the* avoid set *of player i in coalition S as*

$$Av(i, S) = \{S' \subseteq S : (i \in S') \wedge (\forall S'' \subseteq S, S' \preceq_i S'')\}.$$

A game satisfies bottom responsiveness *if for each $i \in N$ and for each pair $S, T \in \mathcal{N}_i$ the following conditions hold:*

(i) *if for each $S' \in Av(i, S)$ and for each $T' \in Av(i, T)$ it holds that if $S' \succ_i T'$, then $S \succ_i T$;*

(ii) *if $Av(i, S) \cap Av(i, T) \neq \emptyset \wedge |S| \geq |T|$, then $S \succeq_i T$.*

In what follows, we assume a minimum a priori knowledge of the values. Namely, we assume to know $v_i(\{i\}), \forall i \in N$. A similar, yet significantly stronger, assumption was used by (Sliwinski and Zick 2017) to prove that Top Responsive HGs (i.e. HGs which satisfy top responsiveness) are efficiently PAC stabilizable.

Anonymous. A HG is said to satisfy *anonimity*, as defined in (Banerjee, Konishi, and Sönmez 2001; Bogomolnaia and Jackson 2002), if $v_i(S) = v_i(T)$ for any player $i \in N$ and any $S, T \in \mathcal{N}_i$ with $|S| = |T|$, i.e., players evaluate coalitions only according to their size.

We are now ready to state the following theorem, summarizing our results for the just defined HGs classes.

Theorem 3. *The results described in Table 1 hold.*

Proof. We give here just a sketch of the proof.

Friends and Enemies. The efficient PAC learnability of both Friends Appreciation and Enemies Aversion profiles follows directly by observing that they are both subclasses of Additively Separable HGs (see (Dimitrov et al. 2006)), known to be efficiently PAC learnable by the results of Sliwinski and Zick (2017).

For what concerns stabilizability, Suzuki and Sung (2010) showed that Friends and Enemies Games under Friends Appreciation are a subclass of Top Responsive HGs. Sliwinski and Zick (2017) proved that Top Responsive HGs are efficiently PAC stabilizable, which then implies the same for Friends Appreciation. For Enemies Aversion, Dimitrov et al. (2006) prove that core-stable partitions always exist, while Dimitrov and Sung (2004) provide an algorithm returning such a partition. Inspired by their algorithm, we provide an algorithm PAC stabilizing this class.

Bottom Responsive. To show that the class is not efficiently PAC learnable, we prove that its pseudo-dimension is lower bounded by $2^{\frac{n-1}{2}}$, and thus is exponential in n. The result then follows by Theorem 1. The construction in our proof bears similarities to the one of Sliwinski and Zick (2017) for Top Responsive HGs.

Regarding stabilizability, we first observe that Suzuki and Sung (2010) show that a core-stable coalition structure always exists for this class. Moreover, a simple necessary condition for S to be part of a core-stable partition π, is that for all $i \in S$ it must hold that $\{i\} \in Av(i, S)$. Indeed, if this condition is not satisfied, at least one player prefers to deviate to a singleton. To give a viable alternative for checking the condition while knowing the values of the singletons, we prove the following property: Given a Bottom Responsive HG $H = (N, v)$, for every $i \in N$ and every $S \in \mathcal{N}_i$, it holds that $\{i\} \in Av(i, S) \iff v_i(\{i\}) \leq v_i(S)$. Starting from this property, we construct Algorithm 1 which, given a sample, returns a coalition structure that is not core-blocked by any coalition from the sample. By Theorem 2 this is sufficient for concluding the efficient PAC stabilizability.

Anonymous. To show efficient PAC learnability, we prove that the pseudo-dimension of this class is upper bounded by $n(1 + \log n)$, and thus is polynomial. Then, for each $i \in N$, the following procedure computes a hypothesis v_i^* consistent with the sample in time polynomial in n and m: For every coalition C of size $k \in [n]$, if there exists S_j s.t. $i \in S_j$ and $|S_j| = k$, then set $v_i(C) = v_i(S_j)$, otherwise set $v_i(C) = -\infty$.

For what concerns stabilizability, we can provide a counter-example showing that the class is not PAC stabilizable, even in the case of natural single-peaked preferences, where every player has a given preferred size, and the valuation decreases as the distance from such size increases. $\square$

Notice that, according to the above theorem, the negative results on stabilizability of Additively Separable HGs of (Sliwinski and Zick 2017) do not transfer to Friends and Enemies Games. Furthermore, while the Bottom Reponsive HGs class is not PAC learnable but efficiently PAC stabilizable, exactly the opposite holds for Anonymous HGs.

Algorithm 1: Stabilizing Bottom Responsive HGs

Input: $N, \mathcal{S} = \{(S_j, \vec{v}(S_j))\}_{j=1}^m$
Output: π: an ε-stable partition of N

1   $\pi \leftarrow \varnothing, \mathcal{T} \leftarrow \varnothing$
2   **for** $\langle S, \vec{v}(S) \rangle \in \mathcal{S}$ **do**
3     $f \leftarrow 1$
4   **if** $\exists i \in S$ *s.t.* $v_i(S) < v_i(\{i\})$ **then** $f \leftarrow 0$ **break**
5   **if** *f=1* **then** $\mathcal{T} \leftarrow \mathcal{T} \cup \{S\}$
6   **while** $\mathcal{T} \neq \varnothing$ **do**
7       $T^+ \leftarrow \operatorname{argmax}_{T \in \mathcal{T}} |T \setminus \bigcup_{P \in \pi} P|$
8       $\pi \leftarrow \pi \cup \left(T^+ \setminus \bigcup_{P \in \pi} P\right)$
9       $\mathcal{T} \leftarrow \mathcal{T} \setminus \{T^+\}$
10      $N \leftarrow N \setminus T^+$
11   **for** $i \in N$ **do** $\pi \leftarrow \pi \cup \{\{i\}\}$
12 **return** π

A General Framework for Learnability: Hedonic Coalition Nets

To provide a general unifying framework for learnability of HGs, a direction worth investigating is the one of determining a suitable superclass or a small number of superclasses encompassing all the learnable HG classes. Such results would contribute to the general understanding of the crucial properties leading to learnability, or lack thereof, and would also provide means to easily determine whether a specific class of HGs is learnable.

A universal HGs class, maintaining the full expressiveness for representing any HG, is the one of the so-called *Hedonic Coalition Nets* (HCNs) (Elkind and Wooldridge 2009). Before giving the definition, we note that, since there exist classes of HGs that are not learnable, it is not possible to get a positive result for the learnability of any fully expressive HGs representation, so not for HCNs either, without imposing further restrictions. Thus, our goal here is to determine suitable restrictions allowing for efficient learnability.

Definition 2. *A* hedonic coalition net (HCN) *is a tuple* $(N, R_1, \dots, R_n)$ *where* N *is a set of* variables *(each corresponding to a player) and* R_i *is the set of* rules *for player* i. *A single rule in* R_i *is given by a pair* (ϕ, β), *where* ϕ *is a formula of propositional logic over* N *and* $\beta \in \mathbb{R}$ *is a real number. We will denote a rule in* R_i *by* $\phi \mapsto_i \beta$. *Then, assuming the conventional semantic satisfaction relation "$\models$", the valuation of player* i *for a coalition* $S \in \mathcal{N}_i$ *is*

$$v_i(S) = \sum_{\substack{\phi^j \mapsto_i \beta^j \in R_i: \\ S \models \phi^j}} \beta^j. \tag{1}$$

The first HCNs subclass we consider comprises HCNs in which the formulas appearing in each set of rules R_i are known a priori. Namely, for any rule $\phi \mapsto_i \beta \in R_i$ we only need to learn β. We first show that, in this case, the pseudo-dimension depends on the number of rules.

Proposition 4. *Let* $\mathcal{H}(R_i)$ *be the class of valuation functions that can be expressed with a fixed set of a priori known distinct rules* R_i. *Then,* $P_{dim}(\mathcal{H}(R_i)) = O(|R_i|)$.

Proof. Let $r = |R_i|$. We will show that no set of size $r + 1$ can be pseudo-shattered by $\mathcal{H}(R_i)$. As a consequence $P_{dim}(\mathcal{H}(R_i)) \leq r$, which implies the result. Let $\mathcal{S} = \{S_j\}_{j=1}^{r+1}$ be any set of coalitions from $\mathcal{N}_i$ of size $r + 1$, and $(t_1, \ldots, t_{r+1})$ any sequence of $r+1$ real numbers. Given any labeling l, the condition $v_i(S_j) > t_j \Leftrightarrow \ell_j = 1$ can be written as a system of $r + 1$ linear inequalities of the form

$$\sum_{k=1}^{r} a_{jk}\beta_k > t_j \text{ if } \ell_j = 1, \text{ and } \sum_{k=1}^{r} a_{jk}\beta_k \leq t_j \text{ if } \ell_j = 0,$$

where $a_{jk} = 1$ if $S_j \models \phi^k$ and 0 otherwise. This is a system of $r + 1$ inequalities with r unknowns $\beta_1, \ldots, \beta_r$, thus the coefficient matrix $A = (a_{jk})$ must have linearly dependent rows. Let us w.l.o.g. assume that the last row A_{r+1} can be written as $A_{r+1} = \sum_{j=1}^{r} y_j A_j$ where the coefficients y_j are not all null. Let us define the labelings $\ell^{(1)}, \ell^{(2)}$ in this way: $\ell_j^{(1)} = 1 \Leftrightarrow y_j < 0$ for $j \in [r]$, $\ell_{r+1}^{(1)} = 1$ and $\ell_j^{(2)} = 0 \Leftrightarrow \ell_j^{(1)} = 1$. By contradiction, assume that there exist solutions $\vec{b}_1$ and $\vec{b}_2$ that satisfy the respective systems of inequalities. Let us consider the first system. By definition of $\ell^{(1)}$ and $\vec{b}_1$, if $\ell_j^{(1)} < 0$ then $\left(A_j \cdot \vec{b}_1\right) > t_j$ but $y_j < 0$ implying that $y_j \left(A_j \cdot \vec{b}_1\right) < y_j t_j$. When $\ell_j^{(1)} = 0$, instead, it holds that $y_j \left(A_j \cdot \vec{b}_1\right) \leq y_j t_j$. We can then conclude that this last inequality holds for all $j \in [r]$. Regarding $\ell^{(2)}$, with the same line of reasoning one can prove that $y_j \left(A_j \cdot \vec{b}_2\right) \geq \sum_{j=1}^{r} y_j t_j$ for all $j \in [r]$. Writing A_{r+1} as a combination of the other rows, and including the inequalities associated to S_{r+1}, we obtain the following:

$$t_{r+1} < A_{r+1} \cdot \vec{b}_1 = \sum_{j=1}^{r} y_j \left(A_j \cdot \vec{b}_1\right) \leq \sum_{j=1}^{r} y_j t_j$$

$$t_{r+1} \geq A_{r+1} \cdot \vec{b}_2 = \sum_{j=1}^{r} y_j \left(A_j \cdot \vec{b}_2\right) \geq \sum_{j=1}^{r} y_j t_j$$

implying $t_{r+1} < \sum_{j=1}^{r} y_j t_j \leq t_{r+1}$, a contradiction. $\square$

We say that $\mathcal{H}$ admits a *compact* HCN representation if it is possible to represent every $v \in \mathcal{H}$ with a polynomial number of rules for each player i. Observe that so far every class that has been shown to be learnable, also admits a compact HCNs representation. Elkind and Wooldridge (2009) give HCNs representations for Additively Separable, Anonymous, $\mathcal{W}$ and $\mathcal{B}$-games.

The following result shows that HGs admitting a compact HCN representation, for which we know the formulas a priori, are efficiently PAC learnable.

Theorem 5. *Let $\mathcal{H}$ be a class of HGs that admits a compact HCN representation. Suppose that for every set of rules R_i, we know the corresponding set of formulas Φ. Then, $\mathcal{H}$ is efficiently PAC learnable.*

While the class presented above includes Additively Separable, Fractional, and Anonymous HGs, which have all

been shown to be efficiently PAC learnable, there exist other learnable classes which do not fall within the above characterization. Indeed, for $\mathcal{W}$-games and $\mathcal{B}$-games, knowing the ϕ for each rule a priori is not possible, since the formulas themselves depend on the ordered preferences that we need to learn. On the other hand, the maximum number of distinct coalition values in both cases is only n.

To capture these remaining classes of learnable HGs through another suitable subclass of HCNs, we resort to *decision lists*, which were introduced by Rivest (1987) as alternative representations for Boolean functions.

Definition 3. *A decision list (DL) L is defined by a set of l rules $L = \{(\kappa_1, b_1), \ldots, (\kappa_l, b_l)\}$ such that κ_i is a conjunction of literals, $b_i \in \{0, 1\}$, $\forall i \in [l]$, and κ_l is the constant function **true**. Given L and a truth assignment x, $L(x)$ is equal to b_j where j is the least index such that $\kappa_j(x) = 1$.*

We use the term k-decision lists (k-DL) if all the conjunctions in the DL are of size at most k.

For our purposes, for any coalition S and $i \in [l]$, $\kappa_i(S) = 1$ if $S \models \kappa_i$ and $\kappa_i(S) = 0$ otherwise. It is convenient to think of a DL as an "if – **then** – **else if** – ... - **else** –" rule.

$$\text{if } \kappa_1(S) = 1 \text{ then return } b_1$$
$$\text{else if } \kappa_2(S) = 1 \text{ then return } b_2$$
$$\cdots$$
$$\text{else return } b_l$$

Note that all the HGs classes we mentioned, other than Anonymous HGs can be represented as k-DL with k constant. It is known that for constant k, k-DL are efficiently PAC learnable: Rivest (1987) shows an efficient learning procedure LEARN-k-DL$(k, \mathcal{S})$ that takes the size k and a sample $\mathcal{S}$ as input and returns a k-DL L. Furthermore, in the same work it is proven that k-DL are strictly more expressive than k-CNF and k-DNF formulas, and decision trees of depth k, meaning that every Boolean function that is representable in one of these forms admits a representation as a k-DL, but not viceversa.

Now, if we consider HCNs which contain rules that are represented by k-DL and additionally restrict our attention to representations in which every coalition satisfies exactly one rule, it turns out that we can again efficiently PAC learn the valuations, as shown in the following.

Theorem 6. *Let $\mathcal{H}$ be a class of HGs that admits a HCN representation such that*

(i) every coalition $S \in \mathcal{N}_i$ satisfies exactly one rule in R_i,
(ii) every rule is of the form $L \mapsto \beta$, where L is a k-DL with k constant, and β unique, i.e., no pair of distinct rules have the same value β.

Then, $\mathcal{H}$ is efficiently PAC learnable by Algorithm 2.

While the second assumption in Theorem 6 seems rather strong, we argue that asking for unique values β actually does not impose a further restriction, even though it seems fundamental for proving the result. Indeed, if there is more than one k-DL associated with the same value β, using the assumption that every coalition satisfies exactly one rule, it is always possible to merge them into one k-DL.

Algorithm 2: Learning HCN in k-DL form

Input: $k \in \mathbb{N}$, $\mathcal{S} = \{(S_j, v_i(S_j))\}_{j=1}^m$
Output: R_i consistent with $\mathcal{S}$
1 $R_i \leftarrow \varnothing$
2 **for** β *in* $\{v_i(S_j) : S_j \in \mathcal{S}\}$ **do**
3   **for** $S_j \in \mathcal{S}$ **do**
4   **if** $v_i(S) = \beta$ **then** $b_j = 1$ **else** $b_j = 0$
5     $\mathcal{S}' = \{(S_j, b_j)\}_{j=1}^m$
6     $L \leftarrow \text{LEARN-}k\text{-DL}(k, \mathcal{S}')$
7     $R_i \leftarrow R_i \cup \{L \mapsto \beta\}$
8 **return** R_i

Theorem 6 includes as a special case all HGs that can be represented by sets of mutually exclusive conjunctions, each containing at most k positive literals. This is so, because we can phrase the negative literals positively within the DL, by associating the presence of such a variable with returning 0. Thus, the conjunction size depends only on the number of positive literals. The case $k = 1$ includes $\mathcal{W}$- and $\mathcal{B}$-games.

Stabilizability of Hedonic Games

We start this section by identifying a property that a HGs class needs to satisfy if it has any ambitions of being PAC stabilizable. To this end, we first define the set of core stable partitions w.r.t. a fixed sample $\mathcal{S}$, and equivalence classes in a HGs class $\mathcal{H}$ w.r.t. a fixed sample $\mathcal{S}$. Then, we state a theorem that abstracts on the arguments used in proofs showing that a specific HGs class is not PAC stabilizable. Recall that a sample $\mathcal{S}$ is a set $\{(S_1, \vec{v}(S_1)), \ldots, (S_m, \vec{v}(S_m))\}$.

Definition 4. *Let $\mathcal{H}$ be a class of HGs, $\mathcal{S}$ a sample and let $H \in \mathcal{H}$. We denote by:*

(i) $C_{\mathcal{S}}(H) = \{\pi : \forall S \in \mathcal{S}, S \text{ does not core block } \pi\}$, *the set of partitions consistent with the sample $\mathcal{S}$;*

(ii) $\mathcal{H}[\mathcal{S}]$ *the set of all instances $H = (N, \vec{v}\,') \in \mathcal{H}$ such that $\vec{v}\,'(S) = \vec{v}(S)$, for each $(S, \vec{v}(S)) \in \mathcal{S}$.*

We are now able to define the following property.

Definition 5. *HGs class $\mathcal{H}$ satisfies the* sample resistant core *property, or has $\mathcal{SRC}$ in short, if for every $\mathcal{S} \subseteq 2^N$*

- $C_{\mathcal{S}}(H) = \emptyset, \forall H \in \mathcal{H}[\mathcal{S}]$, *or*
- $\bigcap_{H \in \mathcal{H}[\mathcal{S}]} C_{\mathcal{S}}(H) \neq \emptyset$.

Theorem 7. *If $\mathcal{H}$ is PAC stabilizable, then $\mathcal{H}$ has $\mathcal{SRC}$.*

Notice that Theorem 7 formalizes the standard approach of Sliwinski and Zick (2017) and of our work, to show that a specific HG class is not PAC stabilizable. Furthermore, property $\mathcal{SRC}$, which is of course satisfied by Top and Bottom Responsive HGs (as they can be PAC stabilized), does not seem to be a common HGs property.

One could argue that aiming to PAC stabilize a specific HG class without having any a priori knowledge on the distribution $\mathcal{D}$ is too ambitious. Thus, a natural question is whether restricting the attention to special distributions increases the prospect of stabilizing some classes of HGs. This

direction was left as an open question by Sliwinski and Zick (2017) and is our focus in the remaining part of the paper.

The motivation for limiting the scope of allowed distributions is to get a more fine-grained insight into PAC stability. The simple counterexamples from (Sliwinski and Zick 2017), while providing valuable understandings, do not reveal "how far away" from achieving PAC stability certain classes of HGs are. Thus, we proceed by studying PAC stability under a class of distributions that excludes the usual adversarial examples. In particular, we focus on distributions having a fair amount of probability mass on all coalitions.

Definition 6. *A distribution $\mathcal{D}$ is said to be* bounded *if there exists $\lambda \geq 1$ such that, for every two coalitions S_1, S_2, it holds that $\Pr_{\mathcal{D}}[S_1] \leq \lambda \Pr_{\mathcal{D}}[S_2]$.*

Observe that the uniform distribution is a special case of the above definition, obtained by setting $\lambda = 1$.

A useful property that we will use extensively in our calculations is that, if $\mathcal{D}$ is bounded with a factor λ, then

$$\frac{1}{\lambda 2^n} \leq \frac{1}{\lambda(2^n - 1)} \leq \Pr_{S \sim \mathcal{D}}[S] \leq \frac{\lambda}{2^n - 1} . \quad (2)$$

These simple bounds follow from the definition of a bounded distribution and the fact that $\sum_{T \in 2^N} \Pr_{\mathcal{D}}[T] = 1$, where the sum goes over the $2^n - 1$ non-empty coalitions. By Equation (2), every coalition now has a positive probability of being sampled. Since the counterexamples to PAC stabilizability of specific HGs classes usually rely on ad hoc distributions where most of the coalitions are never sampled, this feature provides hope of obtaining better results.

$\mathcal{W}$-games under Bounded Distributions

As a case study we consider $\mathcal{W}$-games with no ties. This class admits a polynomial algorithm for finding a core stable partition (Cechlárová and Hajduková 2004), but, despite that, it has been shown not to be PAC stabilizable (Sliwinski and Zick 2017). Thus, it seems a natural first candidate for being PAC stabilizable under bounded distributions. In the rest of this subsection, we indeed show the following result.

Theorem 8. *$\mathcal{W}$-games under bounded distributions are efficiently PAC stabilizable.*

To this end, in what follows, when focusing on a fixed player i, w.l.o.g. we assume that the other players are ordered such that $v_i(1) < v_i(2) < \ldots < v_i(n-1)$. We start by exploiting the fact that the distribution is bounded.

Lemma 9. *Let $\varepsilon > 0$ be fixed. If we denote by $A_j^{(i)}$ the event that a sampled coalition S satisfies $i, j \in S$ and $S \setminus \{i,j\} \subseteq \{j+1, \ldots, n-1\}$, and by $B_j^{(i)}$ the event that a sampled coalition S satisfies $i, j \in S$ and $S \setminus \{i,j\} \subseteq \left\{\lfloor \log_2 \frac{1}{\varepsilon} \rfloor + 2, \ldots, n-1\right\}$, it holds that*

$$\Pr_{S \sim \mathcal{D}}\left[A_j^{(i)}\right] \geq \frac{\varepsilon}{2\lambda} \quad \text{for } 1 \leq j \leq \left\lfloor \log_2 \frac{1}{\varepsilon} \right\rfloor, \text{ and}$$

$$\Pr_{S \sim \mathcal{D}}\left[B_j^{(i)}\right] \geq \frac{\varepsilon}{4\lambda} \quad \text{for } j > \left\lfloor \log_2 \frac{1}{\varepsilon} \right\rfloor.$$

Algorithm 3: Stabilizing $\mathcal{W}$-games

Input: N players, $\mathcal{S} = \{(S_j, \vec{v}(S_j))\}_{j=1}^m, \varepsilon > 0$
Output: A partition π

1   $\vec{v}^* \leftarrow$ LEARN-$\mathcal{W}$-GAMES$(N, \mathcal{S})$
2   $\pi \leftarrow \varnothing$
3   **while** $N \neq \varnothing$ **do**
4   Pick $i \in N$
5   **if** $N \setminus \{i\} \neq \varnothing$ **then**
6         $j \leftarrow \operatorname{argmax}_{k \in N \setminus \{i\}} v_i^*(k)$
7         $\pi \leftarrow \pi \cup \{\{i, j\}\}, N \leftarrow N \setminus \{i, j\}$
8   **else** $\pi \leftarrow \pi \cup \{\{i\}\}$
9   **return** π

Sliwinski and Zick (2017) presented a simple procedure, that we will refer to as LEARN-$\mathcal{W}$-GAMES, which takes in input the set of players and a sample, and returns a consistent estimate $\vec{v}^*$ for the players' valuations in $\mathcal{W}$-games. This procedure sets $v_i^*(j)$ to be the $\max_{S \in \mathcal{S}_{ij}} v_i(S)$ where $\mathcal{S}_{ij} = \{S \in \mathcal{S} : \{i, j\} \subseteq S\}$, if $\mathcal{S}_{ij}$ is non-empty, $-\infty$ otherwise. Next, we define what we call an ε-estimate of a function, and show that the output of LEARN-$\mathcal{W}$-GAMES is actually such an estimate.

Definition 7. *Function v_i' is an ε-estimate of v_i if*

$$\begin{cases} v_i'(j) = v_i(j) & \text{for } 1 \leq j \leq \lfloor \log_2 \frac{1}{\varepsilon} \rfloor, \text{ and} \\ v_i'(j) > v_i(\lfloor \log_2 \frac{1}{\varepsilon} \rfloor) & \text{for } j > \lfloor \log_2 \frac{1}{\varepsilon} \rfloor. \end{cases} \quad (3)$$

Proposition 10. *Let $\varepsilon, \delta > 0$ and $\mathcal{S}$ be a sample of size m. If $m \geq \frac{2\lambda}{\varepsilon} \log \frac{n^2}{\delta}$, LEARN-$\mathcal{W}$-GAMES returns an ε-estimate $\vec{v}^*$ of $\vec{v}$ with confidence $1 - \delta$.*

We want to show that, by relying on the ε-estimate given by LEARN-$\mathcal{W}$-GAMES for ε "not too small", Algorithm 3 returns an ε-stable partition. We first state a technical lemma.

Lemma 11. *Let π be the output of Algorithm 3 and let us call a player i green if it is not in a coalition with one of his $\lfloor \log_2 \frac{1}{\varepsilon} \rfloor$ least preferred choices according to v_i^*. Then,*

a) for i green, $\Pr_{S \sim \mathcal{D}}[i \in S \land v_i(S) > v_i(\pi(i))] < \lambda\varepsilon$,

b) for $\varepsilon \geq \sqrt[3]{\frac{\lambda^2}{2^n}}$, $\Pr_{S \sim \mathcal{D}}[S \text{ does not contain green } i] < \varepsilon$.

We are now finally ready to prove the main theorem, stated in the beginning of this subsection.

Proof of Theorem 8. For $\delta > 0$ and $\varepsilon \geq \sqrt[3]{\frac{\lambda^5}{2^{n-3}}}$, we call Algorithm 3 with $\varepsilon' = \varepsilon/2\lambda$ and $m \geq \frac{2\lambda}{\varepsilon'} \log \frac{n^2}{\delta} = \frac{1}{\varepsilon} \log \frac{n^2}{\delta}$, and obtain an ε-stable partition with probability at least $1 - \delta$. By point a) of Lemma 11, a green node has probability $< \lambda\varepsilon' = \varepsilon/2$ to get a better outcome by moving from $\pi(i)$ to S. Furthermore, ε' satisfies the requirement of point b) in Lemma 11, so the probability of sampling an S without a green node is $< \varepsilon' \leq \varepsilon/2$. In conclusion, if we call G the event that S contains a green player, then since $\Pr_{S \sim \mathcal{D}}[S \text{ core blocks } \pi]$ equals

$$\Pr_{S \sim \mathcal{D}}[S \text{ core blocks } \pi \mid G] + \Pr_{S \sim \mathcal{D}}[S \text{ core blocks } \pi \mid \overline{G}],$$

we see that $\Pr_{S \sim \mathcal{D}}[S \text{ core blocks } \pi] \leq \varepsilon/2 + \varepsilon/2 = \varepsilon$.

For $\delta > 0$ and $\varepsilon < \sqrt[3]{\frac{\lambda^5}{2^{n-3}}}$, with a sample of size $m \geq \frac{8\lambda^6}{\varepsilon^3} \log \frac{n^2}{\delta}$, we can reveal the exact valuation functions with probability at least $1 - \delta$ and return the core stable partition π using the algorithm of Cechlárová and Hajduková (2004). Indeed, since the probability of drawing any coalition is by Equation (2) at least $1/\lambda 2^n$, this also holds for the coalition containing only agents i and j, which provides both $v_i(j)$ and $v_j(i)$. The probability of not drawing a particular coalition of size 2 is $\leq (1 - 1/\lambda 2^n)^m \leq e^{-m/\lambda 2^n} \leq \delta/n^2$. Taking a union bound over all the $\binom{n}{2} < n^2$ coalitions of size 2, we see that the probability of not seeing all the exact valuations is upper bounded by δ. $\square$

Barriers to the Restricted Distributions Approach

Encouraged by the positive results of the last subsection, one could try to extend the approach of focusing on bounded distributions in the hope that other classes that are known not to be PAC stabilizable, such as Additively Separable, Fractional, and Anonymous HGs, are in fact stabilizable under such distributions. Unfortunately, this does not seem to be always the case, as we discuss below.

Definition 8. *For a HGs class $\mathcal{H}$, let $T(\mathcal{H})$ be the time complexity of the best algorithm solving the core for this class, i.e., the runtime of the fastest algorithm that for every input instance either correctly replies that the core is empty or returns a core-stable partition.*

Theorem 12. *If $T(\mathcal{H}) = \omega(poly(2^n))$ for a HG class $\mathcal{H}$, then $\mathcal{H}$ is not efficiently PAC stabilizable, even under the uniform distribution.*

Notice that the assumption $T(\mathcal{H}) \in \omega(poly(2^n))$ is not that strong. In most of the HGs classes the complexity of the problem of deciding the existence of the core is either Σ_2^p-complete or NP-hard. Although this does not imply that it is not possible to find a $O(poly(2^n))$ algorithm (in the case of the total collapse of the polynomial hierarchy, this would even be possible in polynomial time), such algorithms are currently not known and at this point it seems that finding them is unlikely. In particular, the brute force approach, that searches for an element in the core by examining all the possible partitions, has a running time of $\Omega((n/2)^{n/4})$ (as this is one possible lower bound on the Bell number), and thus its running time is also in $\omega(poly(2^n))$.

Conclusions

In this work, we initiated the study of a unified approach for determining the learnability and stabilizability of specific HGs classes. One of the obvious goals for future work is finding a unique characterization of HCN representations that imply learnability. Another one is exploring further consequences of Theorem 12 and expanding the knowledge on the exact computational complexity of solving the core for the different classes of HGs.

Acknowledgements

References

Anthony, M.; and Bartlett, P. L. 2002. *Neural Network Learning - Theoretical Foundations*. Cambridge University Press.

Aziz, H.; and Brandl, F. 2012. Existence of stability in hedonic coalition formation games. In *Proc. 11th Conf. Autonomous Agents and Multi-Agent Systems (AAMAS)*, 763–770.

Aziz, H.; and Savani, R. 2016. Hedonic Games. In *Handbook of Computational Social Choice*, 356–376. Cambridge University Press.

Balcan, M. 2015. Learning Submodular Functions with Applications to Multi-Agent Systems. In *Proc. 14th Conf. Autonomous Agents and Multi-Agent Systems (AAMAS)*, 3.

Balcan, M.; Procaccia, A. D.; and Zick, Y. 2015. Learning Cooperative Games. In Yang, Q.; and Wooldridge, M. J., eds., *Proc. 24th Intl. Joint Conf. Artif. Intell. (IJCAI)*, 475–481.

Balcan, M.; Sandholm, T.; and Vitercik, E. 2018. A General Theory of Sample Complexity for Multi-Item Profit Maximization. In *Proc. 19th Conf. Econom. Comput. (EC)*, 173–174.

Balcan, M.; Vitercik, E.; and White, C. 2016. Learning Combinatorial Functions from Pairwise Comparisons. In *Proc. 29th Conf. Learning Theory (COLT)*, 310–335.

Banerjee, S.; Konishi, H.; and Sönmez, T. 2001. Core in a simple coalition formation game. *Social Choice and Welfare*, 18(1): 135–153.

Bogomolnaia, A.; and Jackson, M. O. 2002. The Stability of Hedonic Coalition Structures. *Games Econom. Behav.*, 38(2): 201–230.

Cechlárová, K.; and Hajduková, J. 2004. Stable partitions with W -preferences. *Discret. Appl. Math.*, 138(3): 333–347.

Dimitrov, D.; Borm, P.; Hendrickx, R.; and Sung, S. C. 2006. Simple Priorities and Core Stability in Hedonic Games. *Social Choice & Welfare*, 26(2): 421–433.

Dimitrov, D.; and Sung, S. C. 2004. Enemies and friends in hedonic games: individual deviations, stability and manipulation. *CentER Discussion Paper Series*.

Dreze, J.; and Greenberg, J. 1980. Hedonic Coalitions: Optimality and Stability. *Econometrica*, 48(4): 987–1003.

Elkind, E.; and Wooldridge, M. J. 2009. Hedonic coalition nets. In *Proc. 8th Conf. Autonomous Agents and Multi-Agent Systems (AAMAS)*, 417–424.

Igarashi, A.; Sliwinski, J.; and Zick, Y. 2019. Forming Probably Stable Communities with Limited Interactions. In *Proc. 33rd Conf. Artificial Intelligence (AAAI)*, 2053–2060.

Jha, T.; and Zick, Y. 2020. A Learning Framework for Distribution-Based Game-Theoretic Solution Concepts. In *Proc. 21st Conf. Econom. Comput. (EC)*, 355–377.

Kearns, M. J.; and Vazirani, U. V. 1994. *An Introduction to Computational Learning Theory*. MIT Press.

Lev, O.; Lu, W.; Tsang, A.; and Zick, Y. 2021a. Learning Cooperative Solution Concepts from Voting Behavior: A Case Study on the Israeli Knesset. In *Proc. 20th Conf. Autonomous Agents and Multi-Agent Systems (AAMAS)*, 1572–1574.

Lev, O.; Patel, N.; Viswanathan, V.; and Zick, Y. 2021b. The Price is (Probably) Right: Learning Market Equilibria from Samples. In *Proc. 20th Conf. Autonomous Agents and Multi-Agent Systems (AAMAS)*, 755–763.

Rivest, R. L. 1987. Learning Decision Lists. *Mach. Learn.*, 2(3): 229–246.

Sliwinski, J.; and Zick, Y. 2017. Learning Hedonic Games. In *Proc. 26th Intl. Joint Conf. Artif. Intell. (IJCAI)*, 2730–2736.

Suzuki, K.; and Sung, S.-C. 2010. Hedonic coalition formation in conservative societies. *Available at SSRN 1700921*.

Trivedi, P.; and Hemachandra, N. 2021. Noise Robust Core-Stable Coalitions of Hedonic Games. arxiv:2109.07738.

Valiant, L. G. 1984. A Theory of the Learnable. *Commun. ACM*, 27(11): 1134–1142.