Adaptive Dynamic Coalition Structure Generation

Lucia Cipolina-Kun
University of Bristol
Bristol, UK
lucia.kun@bristol.ac.uk

Ignacio Carlucho
Heriot-Watt University
Edinburgh, UK

Kalesha Bullard
Google Deepmind
UK

Abstract

We propose a novel framework for the problem of dynamic coalition formation using Deep Reinforcement Learning (DRL). Our approach models the coalition formation process as an ongoing, adaptive deal-or-no-deal game, where each state represents a potential coalition configuration, allowing agents to continually re-assess and alter their affiliations in response to evolving conditions. Notably, our model facilitates the formation of socially-optimal coalitions without requiring the explicit revelation of coalition values among agents. We tested our method in a spatially-distributed ridesharing game. Our results demonstrate robust generalization of the policy to unseen coalitions.

1 Introduction

A dynamic coalition structure generation game (CSG) is a type of strategic interaction where agents can form or dissolve coalitions over multiple time periods. As the game progresses in time, the coalition structure changes due to various external factors, such as new agents entering or exiting the game, or agents changing their location, like in the case of spatial games. Along the game, agents update their information and choose to stay in their current coalition or join new ones. The main elements of this game include a time step, a set of players, feasible coalitions, and a characteristic value function that values formed coalitions . All of which is time dependent. Examples of dynamic games are ridesharing games (Qin et al., 2021), where individuals can form coalitions to achieve economies of scale. Each day, riders can choose to share the ride with different individuals. Notably, these games are coordinated by a central planner such as a ridesharing app, but the ultimate decision to leave or enter a coalition is decided by the rider. Additionally, on game like these, agents do not know the full characteristic function, i.e., the value of the trip for other agents, but only their own cost. Another common application of dynamic coalition games is in the wireless communication domain, typically in MIMO settings (multiple inputs, multiple outputs) where devices in close proximity share a communication channel to enhance data transmission efficiency (Saad et al., 2008; Thi et al., 2017).

Traditional approaches often require exhaustive evaluation of the characteristic function to identify the optimal coalition structure at each time step. This process involves iterating through all possible coalitions, which is computationally intractable due to the exponential number of possible coalition structures and an NP-hard problem. Specifically, for any given characteristic function, the total number of coalition structures is approximately $B_n \sim \theta(n^n)$, where B_n is the n-th Bell number. Moreover, if the game is *dynamic*, agents enter or leave the game changing the original characteristic function. To deal with this complexity, we introduce a Deep Reinforcement Learning (DRL) methodology where agents are trained to learn optimal coalition formation strategies in dynamic games. At inference time, agents are able to identify the socially optimal coalition structure in a single step, thereby bypassing the need for an exhaustive search through the coalition structure space. Additionally, our methodology yields agents capable of adapting to previously unencountered characteristic functions, provided that the game type remains consistent (e.g. spatial games).

2nd Workshop on Agent Learning in Open-Endedness (ALOE) at NeurIPS 2023.

Our first contribution is to propose a DRL method for dynamic CSG, where the environment has been designed to allow the addition or removal of agents after each iteration of the game. In traditional Game Theory, dynamic games have an extra layer of complexity to the already combinatorial nature of CSG games, as each time a new agent enters or exits the game, the characteristic function is effectively altered, leading to a new landscape of coalition values. However, as DRL agents are trained for dynamic environments, they can leverage from their learned policies and can efficiently re-evaluate the optimal coalition structures *in one shot* without the need for exhaustive re-computation of the characteristic function. Additionally, since our methodology is that of single-agent reinforcement learning, we avoid the problem of non-stationarity in multi-agent learning Albrecht et al. (2023).

Our second contribution is to achieve generalization in dynamic spatial games, without agents explicitly observing the characteristic function. In traditional CSG models, a common assumption is that agents posses knowledge of the utility or payoff associated with every possible coalition, as well as observability of the strategies and payoffs of all participating agents. This assumption is often unrealistic in many real-world applications where agents operate under conditions of limited or incomplete information. In contrast, the framework proposed in this paper alleviates the necessity for complete information. Agents iteratively learn optimal strategies for coalition formation based solely on contextual information (such as location, for a spatial game) and their observed rewards. This not only makes the approach more versatile, but it also extends its applicability to a broader range of scenarios where complete game-theoretic information is unattainable or impractical to gather.

2 Background and Related Research

Dynamic coalition games are traditionally studied under cooperative game theory, where there is a centralized authority assigning players into coalitions in order to maximize social welfare (Aumann & Hart, 1994). Yet, in numerous practical scenarios, the reality involves rational agents who form coalitions primarily to further their own agendas (Mahdiraji et al., 2021; Levando, 2017). During the coalition formation process, agents strategically align themselves with coalitions that maximize their own individual utility.

The traditional approach for dynamic coalition structure generation is the *merge-and-split* algorithm applied for example in antenna problems (Saad et al., 2008; Mashayekhy & Grosu, 2011). The algorithm is a deterministic method that operates by iteratively merging and splitting coalitions to find a stable or near-optimal coalition structure based on a utility evaluation. While it can adapt to dynamic changes, the adaptability is often limited to predefined utility functions that have to be evaluated at each time step. In contrast, DRL provides a stochastic, adaptive framework that allows agents to learn optimal coalition-forming strategies over time. While the Merge-and-Split algorithm requires explicit utility functions for its operation, Reinforcement Learning relies on a reward structure, potentially offering greater adaptability in highly dynamic settings. Q-learning was previously used to solve the problem of dynamic coalition formation (Cote et al., 2007). In this work, the coalition formation game was a sequential Markovian decision process in which agents take turns to select a coalition and receive their marginal contribution as rewards. The work's focus is on achieving stable coalition structures that lie in the core. On the other hand, our work focuses on achieving coalitions that maximize the social optimum. Furthermore, while their model benefits from the adaptiveness of RL, it does not include a deep network, which deters its generalization capabilities.

Later works such as the ones from Chalkiadakis & Boutilier (2004); Bachrach et al. (2020) include the DRL component in coalition formation and payoff distribution. The first one is a representative paper on opponent-modelling through Bayesian priors and the second one is a representative paper on bargaining coalitional games. Our approach takes the adaptiveness and generalization of deep networks; however, we are the first to propose a model suited for dynamic spatial games that achieves in-distribution generalization Kirk et al. (2023). Moreover, our method is able to perform well on unseen characteristic functions as long as they are all drawn from the same data distribution.

Our approach diverges from prior studies in its aim to offer a dynamic and adaptive framework. Specifically, we focus on achieving two objectives: first, a coalition selection process for generating socially optimal coalition structures without full information; and second, that is responsive to environmental shifts, such as the entry or leaving of a new agent. For this, we model the coalition

structure generation game as a sequential game where agents take *independent* actions that lead to the social optimum.

3 Problem Formulation

A dynamic coalition formation game in characteristic form is a coalitional game that progresses over time periods, denoted as $T = \{1, 2, \ldots, t\}$ with t finite. It is formally represented by the tuple $\langle (N_t)_{t\in T}, (v_t)_{t\in T}, (\mathcal{CS})_{t\in T}, (\phi_t)_{t\in T}\rangle$, where $N = \{1, 2, \ldots, n\}$ is the set of players; $v_t : 2^n \to \mathbb{R}$ is the characteristic function at time t assigning values $v(S)$ to each coalition $S \subseteq \mathcal{CS}_t$, representing the total worth the coalition can achieve at that time; $\mathcal{CS}_t \subseteq N_t$ is the coalition structure, indicating how agents are grouped into coalitions; and ϕ_t are the transition rules at time t that define how the coalition structure transitions to the next stage based on current structures and player strategies. In our case, the game evolves in time by agents randomly entering or leaving the game. As such, and coalitions within $\mathcal{CS}_t$ may form, dissolve, or change in composition over time.

Our approach is that of an optimization problem in the context of dynamic coalitional games. At each time step, we aim to find an individual allocation rule (i.e. policy) for the socially-optimal partition of agents. This is, assign agents into coalitions in a manner that maximizes the value of the generated coalition structure $\mathcal{CS}_t$. Formally, the social optimum SO is given by: $SO(t) = \arg\max_{\mathcal{CS}_t} \sum_{S\in\mathcal{CS}_t} v(S)$. The objective is to find a coalition structure $\mathcal{CS}_t^*$ that maximize the social optimum, as follows:

$$\mathcal{CS}_t^* = \arg\max_{\mathcal{CS}\in N_t} \sum_{S\in\mathcal{CS}_t} v(S)$$

If this social optimum is unique, and if agents aim to maximize value, it follows that agents would prefer to be part of coalitions that constitute this social optimum. Consequently, the game would reach a point of stability under this value-maximization criterion. Formally, $\forall i \in N, \forall S, D \in SO$ where $i \in S$, $v(S) \geq v(D)$. This condition implies that for each agent i in coalition C, there is no other coalition D in the coalition structure SO that offers a higher total value. Under these conditions, the game reaches a point that we term as *value-maximizing stability*.

In the context of DRL, our objective is to identify an optimal policy $\pi^*(t) : N_t \to \mathcal{S}_t$ that ultimately maps each agent to a coalition within the optimal coalition structure $\mathcal{CS}_t^*$. Formally, the optimization problem can be stated as:

$$\pi^* = \arg\max_{SO} \mathbb{E}\left[\sum_{CS\in SO} v(S)\right] \tag{1}$$

By optimizing this policy, we aim to ensure that agents are allocated into coalitions in such a way that the social optimum is achieved, thereby maximizing the total value generated by these coalitions.

4 Dynamic CSG Problem as a Deal-Or-No-Deal Game

To solve a dynamic CSG we train a single-agent reinforcement learning policy per agent. At each time step, given the observation of the current agents and the coalitions available (and possibly some other contextual data), the goal is for each agent to follow its policy and select the optimal coalition to join such that the social optimum is maximized. At a given time step, after each agent in N_t has made a selection, we obtain a set of *strategy profiles* of coalitions selected by each of the n agents: $(x_1, \ldots, x_n)$. In our protocol, following Slikker (2001) and Kramár et al. (2022), a strategy profile induces a particular $\mathcal{CS}_{\sqcup}$ where agents i and j are on the same coalition iff $x_i = x_j$, i.e. they prefer the same coalition. Additionally, requiring $i \in x_j$ and $j \in x_i$.

The following is the policy training process to obtain an optimal policy per agent. The overall idea is that at each time step, each agent is randomly exposed to different coalitions and contextual data (such as a distance matrix of agents) and receieves an informative reward that guides the actions to the socially optimal choices. The training is sequential over the time domain, but for simplicity, let's fix a time step. At each time step t, the environment randomly determines the number n of agents in play, as such, are 2^n possible coalitions to join. The environment selects an agent $i \in n$ at

random with replacement and the state space $\mathcal{S}_i$ contains all possible coalitions the agent i can be a part of. The observation space for the agent is: (s, s', ϕ_s, ϕ'_s) where s is the agent's current coalition, s' is a different coalition chosen by the environment randomly without replacement from $\mathcal{S}_i$, this is called the *proposed coalition*. Lastly, $\phi_s, \phi_{s'}$ are additional environment variables. For example, in a spatial game, they represent the location coordinates of agents in s and s' . The available actions for an agent are: $\mathcal{A} = \{\text{"Accept", "Reject"}\}$. The reward acts as an informative signal of whether the agent took the correct action in the direction of maximizing the social optimum, and is the only exposure to the characteristic function, albeit indirect. The reward function $R : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ for the accepting action is $R(s, s') = V(s') - V(s)$, and for the rejecting action is $-R(s, s')$. The accepting action transitions the agent from its current coalition to the proposed coalition, i.e. from s to s' and the rejecting action leaves the agent on its current coalition s. The game continues according to the transition rules (changing the number of agents or their locations) to a state in $t + 1$.

For an agent i transitioning between states s and s' over T episodes, the average reward avg_R(s, s') is defined as:

$$\text{avg_R}(s, s') = \frac{1}{T} \sum_{t=1}^{T} R(s_t, s'_t) \tag{2}$$

The methodology is reminiscent of every-visit Monte Carlo RL (Sutton & Barto, 2018), in particular, the value of a state is based on the average return from multiple visits to that state. The optimality criterion stipulates that an agent will prefer coalition (state) s' over s if the episodic average reward is avg_R$(s, s') > 0$. The reward structure encourages agents to reject coalitions of lower value and accept those of higher value, thus achieving the social optimum.

Figure 1 depicts an example sequence of the game for two steps, where agent A has been casually selected on both steps. As an example, consider a sequence where an agent A is offered (and accepted) coalitions in the following order: $A \rightarrow ABC \rightarrow AB \rightarrow AC$. The rewards received by the agent would be (assuming a discount factor of 1): $(ABC - A) + (AB - ABC) + (AC - AB)$. In a subsequent iteration, the agent might be offered coalitions in a different random order: $A \rightarrow AB \rightarrow AC \rightarrow ABC$ with rewards: $(AB - A) + (AC - AB) + (ABC - AC)$. Because coalitions are offered at random, over time, an agent will experience all possible permutations of coalitions. Appendix A contains a worked illustration of the method.

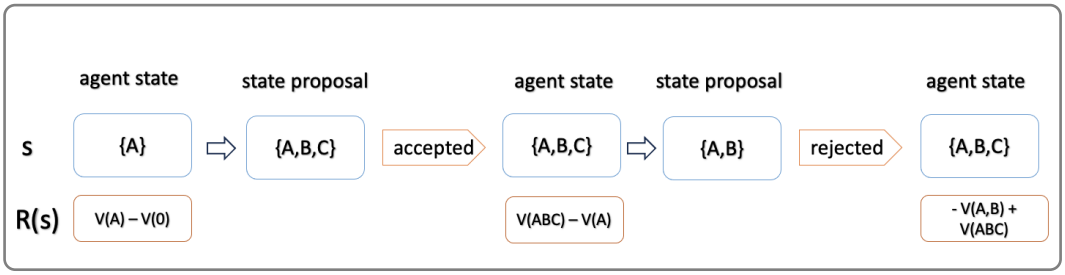


Figure 1: Example of sequence transition for agent A.

5 Experiments

The examples below shows the performance of the methodology on two cost-saving games. We cover the two main types of characteristic functions: subadditive and non-subadditive. In the first type, we expect the grand coalition for form, while on the second, we expect the formation of coalition structures. Additionally, we tested the generalization capabilities of the method to correctly form coalitions in unseen characteristic functions.

5.1 The Ridesharing Game

We introduce a simplified model of a ridesharing game, which is a form of cost game. We assume all agents start from the same point at the origin and travel different distances to the right. The question

Algorithm 1 Reinforcement Learning Methodology for Coalition Formation

Initialize: Coalition set $V = \{A, B, C, \ldots\}$
while Episode not terminated **do**
    Step 1: Select agent A in random fashion
    Step 2: Environment produces state s based on $\mathcal{S}_a$ and other observed variables ϕ
    Step 3: Agent A chooses action: accept or reject proposed coalition s'
    Step 4: Calculate $\Delta V = V(s' \cup \{i\}) - V(s)$
    if Agent ACCEPTS **then**
        Reward $= \Delta V$
    else/ REJECTS
        Reward $= -\Delta V$
    end if
    Step 5: Update agent's coalition (state)
    Step 6: Environment selects the next agent in random fashion
end while
Terminate episode when all permutations of coalitions and agents are depleted

our model answers is whether they can benefit from sharing a ride (i.e. forming coalitions) and if so, with whom. The characteristic function represents the *cost* incurred when agents form a coalition.

$$C(i, S, d, k, \alpha) = \begin{cases} d_i & \text{if } |S| = 1, \\ \frac{k \times \text{var}(d)}{\alpha \times |S|} & \text{otherwise.} \end{cases} \tag{3}$$

Where, d are the distances for each agent i in S. The hyper-parameters κ and α, provide us with the flexibility to test different types of games, as shown below. The parameter k controls the *monotonicity* of the function, $C(S \cup \{i\}) \geq C(S)$. Large values indicate that when adding agents farther away, the increased cost of the trip outweights the benefits. The parameter α controls the *subadditivity* of the function. A game is sub-additive if the cost of merging two coalitions is less than or equal to the sum of their individual costs, i.e., $C(S_1 \cup S_2) \leq C(S_1) + C(S_2)$. In cost games, economies of scale, encourage the formation of larger coalitions, possibly even the grand coalition.

The **rewards** are maximized accounting for a cost game, where agents will form coalition with those who bring less marginal cost. Since $C(\cdot) = -V(\cdot)$ for a current coalition s and a proposed coalition s' we have: $R(s, s') = -C(s') + C(s)$.[1] The **observation space** for each agent is composed of: (s, s') and both distance vectors d_i for $i \in s$ and d_i for $i \in s'$.

5.1.1 Sub additive Game

Low linearity and high convexity. To obtain a subadditive game, we used $k = 1, \alpha = 60$ in equation 3. In subadditive cost games, the economies of scale are strong enough to encourage the formation of the grand coalition. To test the performance of the methodology on these type of games, we trained four different agents, each with its own policy. Figure 2 shows the dynamic formation of coalitions, in particular, all the coalitions accepted during the deal-or-no-deal game. Starting from the left, at t_0, agents start on singletons and as the game progresses to the right, at t_1, Agent B is offered the coalition $\{BD\}$ and accepts it, since in a subadditive game, it brings more value than the singleton. On the same time, agents A and D are offered the coalition $\{ABD\}$, which they accept. According to our protocol, only coalition $\{BD\}$ will form. Agent C doesn't participate again until t_2 when she is offered to join $\{ACD\}$ and accepts. On t_3, $\{ABC\}$ accept to join a coalition together since they are indifferent between coalitions of the same size, while D is not present in the game. Lastly, on t_4 all agents accept to join the grand coalition and on further time steps, every other proposal is rejected (rejected proposals are not shown). To summarize, at each time step, agents are offered several coalitions and accept only those that bring more value to them, achieving the social optimum by virtue.

[1] This reward definition accommodates the cases where the characteristic function contains either all positive or negative terms (but it can't be both at the same time).

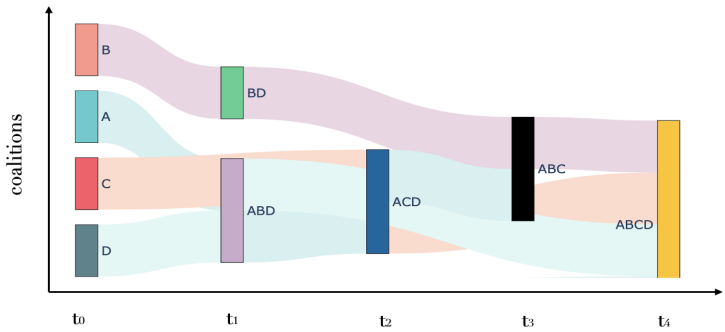


Figure 2: Coalitions accepted by each agent over different time steps. As expected, the game stabilizes in the grand coalition, where every further coalition is rejected.

5.1.2 Experiments on Coalition Structure Generalization

To test the in-distribution generalization of the method, we created samples of tasks (i.e. games) over the domain of spatial games. We created a distribution of 360 games using the game definition in Equation 3 where agent's distances are chosen at random $d_i \in [0.1, 0.5]$ and $k = 20$ to obtain non-subadditive games (i.e. we expect the formation of coalition structures). We trained five agents, each with its own policy as in Equation 1. To measure the performance of each agent's policy, we canlculated the accuracy of the five policies separately. The accuracy metric 4 shows the probability of selecting the correct action given a coalition proposal.

$$\text{Accuracy} = \frac{\text{Number of correct actions}}{\text{Total number of coalitions proposals}} \tag{4}$$

Where *Number of correct actions* sums the number of times the agent has correctly accepted an optimal coalition (i.e. high value) and rejected a sub-optimal coalition (i.e., a coalition with lower value). While the *Total number of coalition proposals* includes the coalitions wrongly accepted or wrongly rejected.

Figure 3 shows the box and whisker plot of the accuracy metric in Equation 4 for each agent against a random policy (dashed line). The green vertical line shows the median. As we can see, all agents perform above the random policy. Different agents show varying degrees of variance, which will be studied in future works.

6 Conclusion

In this work, we presented a DRL method that allows us to form socially-optimal coalitions. We have shown how our methodology has the ability to facilitate optimal decision-making even in the absence of explicit coordination, by just querying from a policy in $O(1)$. Additionally, the stochastic nature of agent and coalition selection by the environment serves as an exploration mechanism, allowing the system to explore various coalition structures until a social optimum is reached. Our method shows that agents learn to select the optimal coalition structure for subadditive and non-subadditive games. Additionally, when equipped with contextual features, the method achieves in-distribution generalization.

7 Future Work

In future works, we plan to study the cause of the different performance metrics for each agent. Some policies exhibiting big variance when compare to others. Additionally, we plan to implement a benchmark model to test against. A natural question in coalition formation is the stability of the formed coalitions. For this, a payoff allocation regime needs to be studied (Magaña & Carreras, 2018). An enhancement in the training step is to improve the generalization power of the deep net.

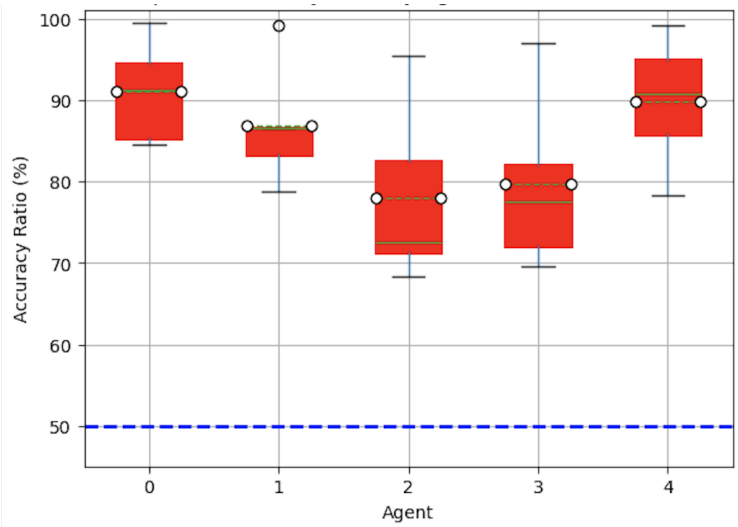


Figure 3: Box plots showing the accuracy of each agent's policy across 360 different game instances. The dashed line represents the accuracy of a random policy. Calculations with one random seed.

For example, through specialized methodologies, such as regularization, to improve the problem of catastrophic forgetting (Kirkpatrick et al., 2017).

Acknowledgments

This work was partly supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through a Turing AI Fellowship (EP/V022067/1) on Citizen-Centric AI Systems (https://ccais.ac.uk/).

References

Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2023.

R. J. Aumann and S. Hart (eds.). *Handbook of Game Theory, with Economic Applications*, volume 2. North-Holland, Amsterdam, 1994.

Y. Bachrach, R. Everett, E. Hughes, A. Lazaridou, J. Z. Leibo, M. Lanctot, M. Johanson, W. M. Czarnecki, and T. Graepel. Negotiating team formation using deep reinforcement learning. *arXiv preprint arXiv:2010.10380*, October 2020.

G. Chalkiadakis and C. Boutilier. Bayesian reinforcement learning for coalition formation under uncertainty. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, pp. 1090–1097, July 2004.

Enrique Munoz de Cote, Alessandro Lazaric, and Marcello Restelli. A learning approach to dynamic coalition formation. *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, 2007. URL `http://researchers.lille.inria.fr/~lazaric/Webpage/PublicationsByType_files/munoz2007learning.pdf`.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76: 201–264, 2023. URL `https://dl.acm.org/doi/epdf/10.1613/jair.1.14174`.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

János Kramár, Tom Eccles, Ian Gemp, Andrea Tacchetti, Kevin McKee, Mateusz Malinowski, Thore Graepel, and Yoram Bachrach. Negotiation and honesty in artificial intelligence methods for the board game of diplomacy. *Nature Communications*, December 2022.

Dmitry Levando. Formation of coalition structures as a non-cooperative game. *arXiv preprint arXiv:1702.06922*, 2017. URL `https://arxiv.org/abs/1702.06922`.

Antonio Magaña and Francesc Carreras. Coalition formation and stability. *Group Decision and Negotiation*, 27:467–502, 2018.

Hannan Amoozad Mahdiraji, Elham Razghandi, and Adel Hatami-Mar. Overlapping coalition formation in game theory: A state-of-the-art review. *Expert Systems With Applications*, 174: 114752, 2021.

Lena Mashayekhy and Daniel Grosu. A merge-and-split mechanism for dynamic virtual organization formation in grids. In *30th IEEE International Performance Computing and Communications Conference*, pp. 1–8. IEEE, 2011. URL `https://doi.org/10.1109/PCCC.2011.6108108`.

Zhiwei Qin, Hongtu Zhu, and Jieping Ye. Reinforcement learning for ridesharing: A survey. *arXiv preprint arXiv:2105.01099*, 2021.

W. Saad, Z. Han, M. Debbah, and A. Hjorungnes. A distributed merge and split algorithm for fair cooperation in wireless networks. In *ICC Workshops - 2008 IEEE International Conference on Communications Workshops*, pp. 311–315, 2008. doi: 10.1109/ICCW.2008.65.

Marco Slikker. Coalition formation and potential games. *Games and Economic Behavior*, 37: 436–448, 2001. doi: 10.1006/game.2001.0846.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Minh-Thuyen Thi, Thong Huynh, Amr Radwan, et al. Coalition formation game for joint power control and fair channel allocation in device-to-device communications underlaying cellular networks. *Wireless Personal Communications*, 96:1173–1191, 2017. doi: 10.1007/s11277-017-4230-3.

A Appendix

A.0.1 Illustration of the method.

Consider for example a game with two agents $\Gamma : (N = \{A, B\}, V(\cdot))$. The goal is for each agent to choose the most valuable coalition (or singleton) without having full information on the characteristic function of a game. On the next step, as a possible dynamics of the game, B leaves and enters C defining a new game Γ between A and C. Again A and C choose the most valuable coalition available. And so on.

To train this type of game, at each step, the environment selects one member at random with replacement, say A, its set of feasible coalitions is: $\mathcal{S}_a = \{\{A\}, \{A, B, C\}, \{A, B\}, \{A, C\}\}$. Denote $s \in \mathcal{S}_a$ the current state of A. The environment proposes a coalition s' chosen at random, without replacement, from $\mathcal{S}_a$. The actions available to A are either to accept or reject the proposal. After the agent's action, the environment records the state of the agent (either s in case of rejection or s' in case of acceptance) and assigns the agent a reward following the Section 4. Next, the game transitions to a new step where *another agent* is selected with replacement and states are selected from the agent's remaining set. The episode terminates when all the permutations of coalitions and agents have been depleted.

B Efficacy of Marginal Contribution Heuristic

We present an example of the application of the method for non-superadditive coalitions with a dominant coalition. The methodology benefits from the fact that the sum of the marginal contributions of members of a coalition equal the total value of a coalition (efficiency).

B.1 Characteristic Function

The characteristic function V for this game is non-superadditive. Leading to the formation of a coalition structure.

$$\begin{aligned} V(\{A\}), V(\{B\}), V(\{C\}) &= 0 \\ V(\{A, B\}) &= 3 \\ V(\{A, B, C\}) &= 2 \end{aligned}$$

B.2 Analysis of the Marginal Contributions

Marginal contribution of A and B in $\{A, B\}$ is 3, while in $\{A, B, C\}$, it decreases. This high marginal contribution for both A and B in $\{A, B\}$ identifies it as a strong candidate for the core.

In the given characteristic function, our heuristic leverages marginal contributions to spotlight high-value coalitions. For the coalition $\{A, B\}$, the marginal contributions of A and B are both 3, which stands as the highest value across all possible coalitions in this game.

To contrast, let's examine the marginal contributions for the coalition $\{A, B, C\}$:

$$\begin{aligned} \text{MC}(A, \{A, B, C\}) &= V(\{A, B, C\}) - V(\{B, C\}) = 2 - 0 = 2, \\ \text{MC}(B, \{A, B, C\}) &= V(\{A, B, C\}) - V(\{A, C\}) = 2 - 0 = 2, \\ \text{MC}(C, \{A, B, C\}) &= V(\{A, B, C\}) - V(\{A, B\}) = 2 - 3 = -1. \end{aligned}$$

The marginal contributions for the agents in $\{A, B, C\}$ are inferior to those in $\{A, B\}$. Specifically, the marginal contribution of C is negative, thereby indicating that the inclusion of C into the coalition $\{A, B\}$ results in a decrement in the coalition's value. In this case,as per the methodology, when C is offered to transition to coalition $\{ABC\}$ it will receive a negative reward and thus end discarding it. Thus, in line with marginal contributions, our heuristic identifies $\{A, B\}$ as the high-value coalition, effectively disregarding $\{A, B, C\}$.

B.3 Special Characteristics of the Example Function

Why the method works? Is it particular to this example? The answer is that the method is applicable to the most common types of characteristic functions. Namely:

1. **Existence of a Dominant Coalition**: The coalition $\{A, B\}$ has a value of 3, surpassing the value of any other coalition or partition of coalitions. This establishes it as a dominant coalition within the game.
2. **High Marginal Contributions in Dominant Coalition**: In the coalition $\{A, B\}$, the marginal contributions of both A and B are 3. These are the highest marginal contributions across all possible coalitions, making them congruent with our heuristic of using marginal contributions to identify high-value coalitions.
3. **Negative Marginal Contribution for Non-Dominant Additions**: The marginal contribution of C when added to $\{A, B\}$ is negative. This effectively indicates that C should not be included in the dominant coalition, thereby aiding in the facile identification and exclusion of less valuable extensions to the dominant coalition.

Owing to these attributes, our heuristic can adeptly identify $\{A, B\}$ as the dominant coalition, rendering it a likely candidate for a coalition within the game's core.

B.4 Methodological Equivalence to Coalition Values

The underlying substance of the methodology is that rewarding agents by the difference in coalition values over multiple episodes indirectly allows for the comparison of the relative values of coalitions. Importantly, this is achieved without ever disclosing the actual value of any coalition. A motivating example can be found on Appendix A. For each agent i, they go through a series of transitions (s, s') such that i is a member of both s and s'. These coalitions are randomly sampled from the set of all possible coalitions that include i, where $T =$ Permutations involving i. The average difference for an agent i transitioning from s to s' is represented as:

$$\text{avg_R}(s, s') = \frac{1}{T} \sum_{(s,s')} (V(s'_t) - V(s_t)) \tag{5}$$

Where $V(s)$ represents the value of the characteristic function for coalition s. We show that using this average difference as reward acts as an effective heuristic for the agent to gauge the relative value of the coalitions.

Let $\mathcal{S}_i$ be the set of all coalitions that include agent i, randomly ordered. For each coalition $s \in \mathcal{S}_i$, let avg_R(s, s') be the average reward for agent i when it is part of any randomly selected coalition s and transitions to coalition s':

$$\text{avg_R}(s, s') = \frac{1}{T} \left(T \cdot V(s') - \sum_{s \in \mathcal{S}_i, s' \neq s} V(s') \right) \tag{6}$$

Proof:

1. *Comparing Average Rewards Across States*: An agent decides which coalition to join by comparing the average reward of each state. When comparing avg_R(s, s') and any other transition state avg_R(s', s), we get:

$$\frac{1}{T} \left(T \cdot V(s) - \sum_{s' \in \mathcal{S}_i, s' \neq s} V(s') \right) \text{ vs } \frac{1}{T} \left(T \cdot V(s') - \sum_{s \in \mathcal{S}_i, s \neq s'} V(s) \right) \tag{7}$$

After canceling out the $\frac{1}{T}$ terms and the common coalitions in the sums, we are left with:

$$(T + 1) \cdot V(s) \text{ vs } (T + 1) \cdot V(s')$$

This is equivalent to comparing $V(s)$ and $V(s')$ directly, thereby allowing agents to prefer higher-value coalitions without disclosing the actual value of any coalition.

The methodology assigns the marginal contribution $MC(i, S) = v(S \cup \{i\}) - v(S)$ of agent i to coalition S, as a *reward* (and not as the imputation or payoff) to *any agent* entering the coalition on *any order* it acts as a bootstrapping way of approximating $V(S) = \sum_{i \in S} \text{MC}(i, S)$.

C Simple Examples

This section shows the algorithm's capacity to identify the right coalition for the agents on simple games.

C.1 Simple Additive Game

We present a superadditive characteristic function: $V(x) = \sum(x)^3$, where x is the number of agents in the coalition. The aim is for trained agents to be able to identify the coalition with more value for them to join, without seeing the characteristic function. The testing consisted on offering the agents different coalitions and collecting the agent's responses on whether they would like to leave their current coalition and join the offered one. The Figure 4 shows the coalitions offered and the agent's responses. On the left, we see the agents started on their singleton, when moving to the right, we see the sequence of coalitions accepted. Since the game is superadditive, agents accept to join coalitions with more members, and after the grand coalition has been formed, even when more coalitions are offered to agents, they reject the proposal, leading to the stable point of the grand coalition where no agent wants to deviate from.

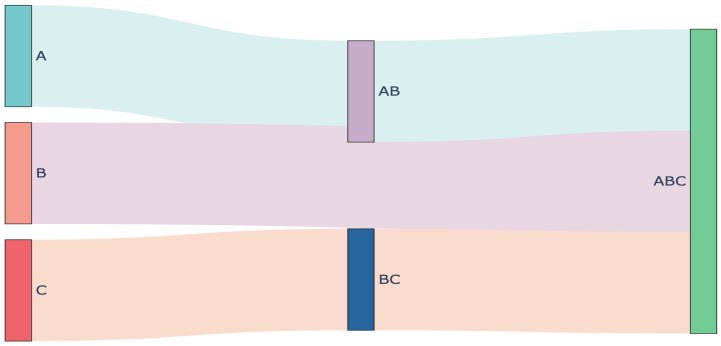


Figure 4: Example of agents playing the deal or no deal game. As expected, the grand coalition is formed.

C.2 Simple Non-additive Game

These examples show how the method is able to form coalition structures for simple non-superadditive games. Consider the following characteristic function for a non-superadditive game.

$$
\begin{aligned}
V(\{A\}), V(\{B\}), V(\{C\}) &= 0 \\
V(\{A, B\}) &= 3 \\
V(\{A, B, C\}) &= 2
\end{aligned}
$$

Figure 5 shows the dynamics of the game. On the left, the game starts with each agent as singleton (each color is a different agent). As the game progresses, to the right, we see the strategy of A and B is to always accept AB, while C's strategy is to merge and join ABC as it is the most beneficial coalition for it. However, once AB form a coalition, they have no incentive to deviate from it and accept ABC as a coalition. The right-most colors show the two final coalitions, composing the coalition structure AB, C

C.2.1 Coalition Structure Generation in the Ridesharing Game

High linearity and low convexity. We have implemented Equation 3 in the case where the cost of adding more agents to a coalition will likely outweigh the benefits, leading to the formation of

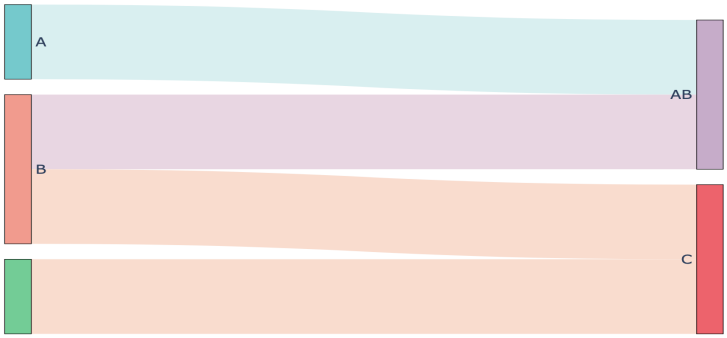
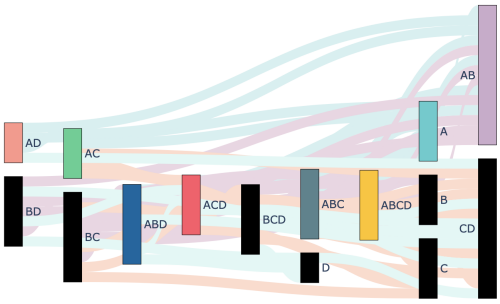


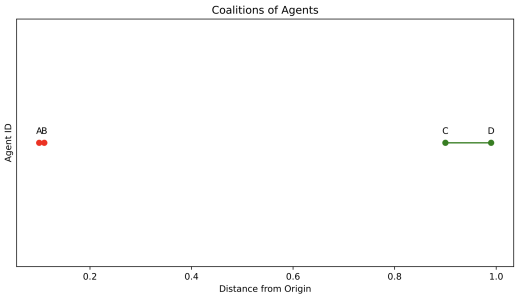
Figure 5: Example of agents playing the deal or no deal game. As expected, C accepts to join the grand coalition but A and B rejects it.

coalition structures. In the limit, for high enough values of k we would only obtain singletons. To test the performance on non-subadditive games, we have used $k = 7$, in a four agent game as shown in Figure 6. On the left pane, starting from the left, we see the coalitions formed by each agent until a stability point is reached in *accepting* $\{\{AB\}, \{CD\}\}$, from there on wards, every further coalition is rejected. Panel (b) shows the location of agents in the spatial game, which validates the final formation of the coalition structure.



(a) Dynamics of a coalition structure formation game. Agents reject every coalition until stabilizing on $\{AB\}, \{C, D\}$

.



(b) Final coalition structure against distance

Figure 6: Example of a non-subadditive ridesharing game for four agents.