Rethinking Formal Models of Partially Observable Multiagent Decision Making (Extended Abstract)*

Vojtěch Kovařík[1,2] , **Martin Schmid**[3,4] , **Neil Burch**[3,4] ,
Michael Bowling[3,4] and **Viliam Lisý**[1]
[1]AI Center, FEE, Czech Technical University in Prague
[2]Foundations of Cooperative AI Lab, Carnegie Mellon University, Pittsburgh
[3]University of Alberta, Computing Science, Faculty of Science
[4]DeepMind, Alberta, Edmonton
vojta.kovarik@gmail.com, lisyvili@fel.cvut.cz

Abstract

Multiagent decision-making in partially observable environments is usually modelled as either an extensive-form game (EFG) in game theory or a partially observable stochastic game (POSG) in multiagent reinforcement learning (MARL). One issue with the current situation is that while most practical problems can be modelled in both formalisms, the relationship of the two models is unclear, which hinders the transfer of ideas between the two communities. A second issue is that while EFGs have recently seen significant algorithmic progress, their classical formalization is unsuitable for efficient presentation of the underlying ideas, such as those around decomposition.

To solve the first issue, we introduce factored-observation stochastic games (FOSGs), a minor modification of the POSG formalism which distinguishes between private and public observation and thereby greatly simplifies decomposition. To remedy the second issue, we show that FOSGs and POSGs are naturally connected to EFGs: by "unrolling" a FOSG into its tree form, we obtain an EFG. Conversely, any perfect-recall timeable EFG corresponds to some underlying FOSG in this manner. Moreover, this relationship justifies several minor modifications to the classical EFG formalization that recently appeared as an implicit response to the model's issues with decomposition. Finally, we illustrate the transfer of ideas by presenting three key EFG techniques – counterfactual regret minimization, sequence form, and decomposition – in the FOSG framework.

1 Introduction

Sequential decision-making is one of the core topics of artificial intelligence research. The ability of an AI agent to perform actions, observe their consequences, and perform further actions towards a goal is instrumental in domains from robotics and autonomous driving to medical decision diagnosis and automated personal assistants. Recent progress has led to unprecedented results in many large-scale problems of this type. While conceptually simpler problems can be modelled with perfect information or by regarding the other agents as stationary parts of the environment, realistic models of real-world situations require rigorous treatment of imperfect information and multiple independent decision-makers operating in a shared environment. The most popular game-theoretical model in these setting – extensive form games (EFG) – dates back to 1953 [von Neumann and Morgenstern, 1953]. EFGs have served the community well, and many impressive results build on top of this particular framework [Moravčík *et al.*, 2017; Brown and Sandholm, 2017b; Brown and Sandholm, 2019].

However, EFGs lose crucial information inherently present in many environments: the observations received by the agents. Observations are essential not only for specifying *what* information was received, but also to know *who* received it and *when*. Since EFGs simply group states indistinguishable to the acting player, the notion of information being public or private is forever lost, and so are the data about the timing. However, these concepts are essential for recent search algorithms [Moravčík *et al.*, 2017; Brown and Sandholm, 2017a; Brown *et al.*, 2020], where decomposition and reasoning about subgames crucially rely on the notion of public information. While it is common to try to recover the necessary information from the EFG model [Burch *et al.*, 2014; Johanson *et al.*, 2011; Šustr *et al.*, 2019], we show that this is impossible to do in general. Practical implementations thus bypass the model by using algorithms that are built with game-specific concepts (e.g., dealing cards in poker), rather than developing algorithms running purely on top of the information provided by the formal model.

On the high-level, the key contributions of this paper are as follows: First, we argue that *to get the most value out of game-theoretical models, we should no longer discard information about whether observations in an environment are observed jointly or privately.* We propose that this can be done by using factored-observation stochastic games (FOSGs), a minor extension to the existing partially-observable stochastic game model. Additionally, we show EFGs and POSGs/FOSGs

* This is an extended abstract for [Kovařík *et al.*, 2022].

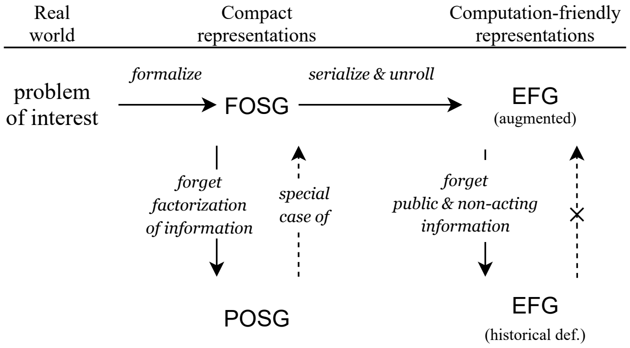

Figure 1: The relationship between Factored-Observation Stochastic Games, Partially-Observable Stochastic Games, and Extensive-Form Games. All (timeable perfect-recall) EFGs can be obtained by starting with some FOSG (Thm. 3.10). Going from augmented to "classical" EFGs is a lossy process that cannot, in general, be reversed (Sec. 3.3)

should not be viewed as two unrelated models. Instead, *EFGs are derived objects that can be obtained by "unrolling" some underlying FOSG.* This both highlights a new application for standalone EFG research and suggests that explicitly reasoning about the underlying FOSG can have significant benefits. (Indeed, FOSGs have the potential to enable new techniques and insights, make results accessible to a broader audience, and align our formal language with recent domain-implementations which often already resemble FOSGs.) More specifically, the full paper shows:

1. There is an an equivalence between POSGs and FOSGs. Any POSG result directly applies to FOSGs as well, since it suffices to merely forget the factorization of observations (Proposition 2.3).
2. We provide a mapping between EFGs and FOSGs. Every FOSG has a canonical extensive-form representation, and every "well-behaved" (perfect recall and timeable) EFG corresponds to some FOSG (Theorems 3.6 and 3.10).
3. Moreover, this relationship between FOSGs and EFGs suggests that our extended definition (Def. 3.5) is a natural way of formalizing EFGs. As a combination of several recent extensions of the EFG model [Burch *et al.*, 2014; Johanson *et al.*, 2011; Brown *et al.*, 2018; Seitz *et al.*, 2019], our formalization removes the need for implementing modern search algorithms in a domain-specific manner. In particular, it preempts various problems with decomposition that are difficult or even impossible to solve in the historical formalization of EFGs (Section 3.3).
4. Finally, we demonstrate that translating EFG results to FOSGs is straightforward: Decomposition in FOSGs has intuitive properties (Section 4.3) and two key EFG techniques – counterfactual regret minimization and sequence form – are easy to formulate in this framework (Sections 4.2 and 4.4).

In the remainder of this extended abstract, we give a formal definition of the proposed model. For a demonstration of the formalism on the example of Kuhn poker, and further details and results, we invite the reader to read [Kovařík *et al.*, 2022].

2 Factored-Observation Stochastic Games

In this section, we describe the factored-observation stochastic game model as a variation on partially-observable stochastic games. Since a key feature of the model is its ability to talk about public information, let us first explain what this concept refers to, why is it important, and why incorporating it into our models is natural, useful, and inexpensive.

2.1 Public Information and Decomposition

A piece of information is said to be **common knowledge** among a group of agents if all the agents know it, they all know that they know it, they all know that they all know that they know it, and so on [Fagin *et al.*, 2003]. Figuring out which information is common knowledge is often difficult, as it requires putting oneself in shoes of the other agents and accurately reasoning about their thought processes. In contrast, a piece of information is **public** among a group of agents if each agent received it in a way that trivially reveals that the information is common knowledge. Some actions that create public knowledge are (a) speaking out loud in a group (b) placing a card face-up, or (c) moving a piece on a game board.

To see why public information is important, first note that the essence of game theory is that playing well requires knowing which actions the other players might take, which tends to involve reasoning about the information they have. Knowing that some information I is *common knowledge* is therefore immensely useful for decomposition. Indeed, each player will only reason about situations compatible with I (and about others' reasoning about situations compatible with I, others' reasoning about reasoning about situations compatible with I, etc.). Situations compatible with I can, therefore, be considered mostly independently of those incompatible with it (the limited dependence is explained in, e.g., [Moravčík *et al.*, 2017]). However, finding out which information is common knowledge can be costly, so the best approach will often be to decompose games based on the subset of common knowledge that is *public*. Once we can decompose problems into smaller pieces, we become able to solve large problems that would otherwise be intractable. Indeed, we have recently seen breakthroughs in a number of problems that were made computationally feasible by decomposition based on public knowledge — examples include poker [Moravčík *et al.*, 2017; Brown and Sandholm, 2017b; Brown and Sandholm, 2019], Hanabi [Foerster *et al.*, 2019; Lerer *et al.*, 2020], general EFGs [Šustr *et al.*, 2019; Li *et al.*, 2020; Davis *et al.*, 2019], Dec-POMDPs [Spaan *et al.*, 2008; de Witt *et al.*, 2019], and one-sided POSGs [Horák and Bošanský, 2019].

However, the existing game-theoretic and MARL models do not have a built-in way of describing public knowledge. This is a serious problem because figuring out whether a piece of information is public or not might be difficult (or impossible) unless one remembers how the information was obtained. (Indeed, suppose I know that a friend would be happy to pick me up from an airport, and they know I would be glad if they did. This alone says nothing about whether the information is public among us — therefore, unless we explicitly talked about this, I should probably take a taxi. For more

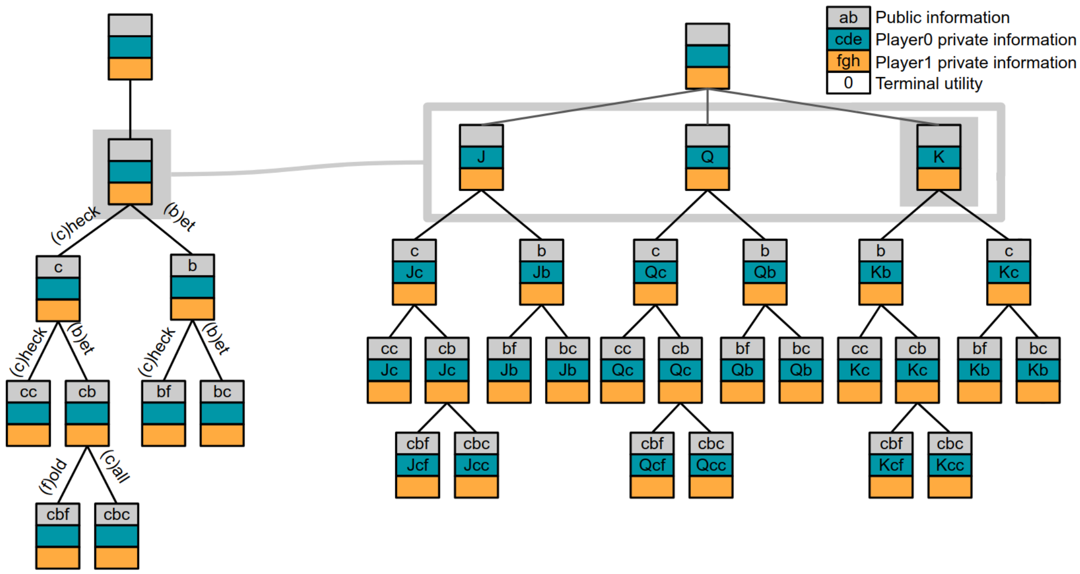

Figure 2: The different points of view enabled by the proposed FOSG model. Bottom (*only shown in the full version of the paper*): the full extensive form representation of Kuhn poker (where the only cards are J, Q, and K). While the tree structure is similar to a classical EFG representation, we retain the factorization of information and the information of the non-acting player. (*Top*) right: the Player 1 point of view, obtained by grouping histories based on which action-observation sequence they correspond to. (*Top*) left: the public tree, obtained by grouping action-observation sequences based on which public states they correspond to.

technical examples, see Section 3.3.) Making use of decomposition thus typically required adding some component on top of the employed formal model. In addition to requiring a non-trivial conceptual and technical effort, this was often done in a domain-specific manner [Moravčík *et al.*, 2017; Brown and Sandholm, 2019], which made the devised methods difficult to generalize. If the generalization were straightforward, the situation in imperfect-information games would by now likely be closer to that in perfect-information games or single-agent RL, where many of the state-of-the-art algorithms are very general [Silver *et al.*, 2017; Schrittwieser *et al.*, 2019; Badia *et al.*, 2020]. This suggests that using a model which keeps track of public information by default would have significant benefits.

Fortunately, as witnessed by the examples (a), (b), and (c), the information about which knowledge is public is inherently a part of the description of many games and real-world situations. In other words, public knowledge is typically *already* a part of a problem's natural definition; the model we propose merely *preserves* this information while *models used in the past discard it*.

2.2 Description of the Basic Model

Informally, the model we are about to describe captures a situation where multiple actors take actions – possibly simultaneously – which influence how the world's state changes. The new world-state might be more or less desirable for individual actors, which is captured by corresponding reward functions. Rather than having full knowledge of the world state, the agents perceive it through observations. These are further "factored", such that if some information is public, each agent will know that everybody else also has access to it. In the following paragraph, $\hookrightarrow$ indicates a *partial function*, defined on a subset of the stated domain.

A **factored-observation stochastic game** (FOSG) is a tuple $G = \langle \mathcal{N}, \mathcal{W}, p, w^{\text{init}}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{O} \rangle$, where $\mathcal{N}$ is the **player set**, $\mathcal{W}$ is the set of **world states**, w^{init} is a designated **initial state**, $p : \mathcal{W} \to 2^{\mathcal{N}}$ is a **player function**, $\mathcal{A}$ is the space of **joint actions**, $\mathcal{T} : \mathcal{W} \times \mathcal{A} \hookrightarrow \Delta\mathcal{W}$ is the **transition function**, $\mathcal{R} : \mathcal{W} \times \mathcal{A} \hookrightarrow \mathbb{R}^{\mathcal{N}}$ is the **reward function**, $\mathcal{O} : \mathcal{W} \times \mathcal{A} \times \mathcal{W} \hookrightarrow \mathbb{O}$ is the **observation function**, and we have:

- $\mathcal{N} = \{1, \ldots, N\}$ for some $N \in \mathbb{N}$.
- $\mathcal{W}$ is compact. For formal convenience, we assume that $p(w^{\text{init}}) = \emptyset$.
- $\mathcal{A} = \prod_{i \in \mathcal{N}} \mathcal{A}_i$, where each $\mathcal{A}_i$ is an arbitrary set of ***i*'s actions**.
  - For each $i \in p(w)$, $\mathcal{A}_i(w) \subset \mathcal{A}_i$ denotes a non-empty compact set of i's (legal) **actions at w**. We denote $\mathcal{A}(w) := \prod_{i \in p(w)} \mathcal{A}_i(w)$.
  - We denote $\mathcal{A}_i(w) := \{noop\}$ for $i \notin p(w)$, which allows us to identify each $a \in \mathcal{A}(w)$ with an element of $\prod_{i \in \mathcal{N}} A_i(w)$ by appending to it the appropriate number of *noop* actions.
- The transition probabilities $\mathcal{T}(w, a) \in \Delta\mathcal{W}$, $a \in \mathcal{A}(w)$, are defined for all $w \in \mathcal{W}$ with non-empty $p(w)$ and for some w with no active players.
  - A world state with $p(w) = \emptyset$ and undefined $\mathcal{T}(w, a)$ is called **terminal**.
- $\mathcal{R}(w, a) = (\mathcal{R}_i(w, a))_{i \in \mathcal{N}}$ for each non-terminal state w and $a \in \mathcal{A}(w)$.
- $\mathcal{O}$ is factored into **private observations** and **public observations** as $\mathcal{O} = (\mathcal{O}_{\text{priv}(1)}, \ldots, \mathcal{O}_{\text{priv}(N)}, \mathcal{O}_{\text{pub}})$.
  - $\mathbb{O} = \prod_{i \in \mathcal{N}} \mathbb{O}_{\text{priv}(i)} \times \mathbb{O}_{\text{pub}}$, where $\mathbb{O}_{(\cdot)}$ are arbitrary

sets (of possible observations).

- We assume that $\mathcal{O}_{(\cdot)}(w, a, w') \in \mathbb{O}_{(\cdot)}$ is defined for every non-terminal w, $a \in \mathcal{A}(w)$, and w' from the support of $\mathcal{T}(w, a)$.

The game proceeds as follows: It starts in the initial state w^{init}. In each state, each active player $i \in p(w)$ learns which actions are legal for them (either by deduction or by being explicitly told) and selects $a_i \in \mathcal{A}_i(w)$. (While the player obviously knows which action *they* took, they often will not know the actions of the other players, or possibly not even which of them were active at w.) The game then transitions to a new state w', drawn from the distribution $\mathcal{T}(w, a)$ that corresponds to the joint action $a = (a_i)_{i \in p(w)}$. This generates the observation $\mathcal{O}(w, a, w')$, from which each player receives $\mathcal{O}_i(w, a, w') := \big(\mathcal{O}_{\text{priv}(i)}(w, a, w'), \mathcal{O}_{\text{pub}}(w, a, w')\big) \in \mathbb{O}_i$ (i.e., the public observation together with their private observation, in manner which allows distinguishing between the two). Finally, each player is assigned the reward $\mathcal{R}_i(w, a)$. However, a player might not know how much reward they received, unless this is – explicitly or implicitly – a part of $\mathcal{O}_i$. This process repeats until a terminal state is reached. The goal of each player is to maximize the sum of rewards $\mathcal{R}_i(w, a, w')$ obtained during the game.

Acknowledgements

We are grateful to Tomáš Gavenčiak, Karel Horák, and Marc Lanctot for discussions related to an earlier version of this paper. This work was supported by the Czech Science Foundation grants no. 18-27483Y and no. GA22-26655S, and the Czech Ministry of Education Youth and Sports grant RCI no. CZ.02.1.01/0.0/0.0/16019/0000765.

References

[Badia *et al.*, 2020] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. *arXiv preprint arXiv:2003.13350*, 2020.

[Brown and Sandholm, 2017a] Noam Brown and Tuomas Sandholm. Safe and nested subgame solving for imperfect-information games. In *Advances in Neural Information Processing Systems*, pages 689–699, 2017.

[Brown and Sandholm, 2017b] Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, page eaao1733, 2017.

[Brown and Sandholm, 2019] Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456):885–890, 2019.

[Brown *et al.*, 2018] Noam Brown, Tuomas Sandholm, and Brandon Amos. Depth-limited solving for imperfect-information games. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7674–7685, 2018.

[Brown *et al.*, 2020] Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. *arXiv preprint arXiv:2007.13544*, 2020.

[Burch *et al.*, 2014] Neil Burch, Michael Johanson, and Michael Bowling. Solving imperfect information games using decomposition. In *AAAI*, pages 602–608, 2014.

[Davis *et al.*, 2019] Trevor Davis, Martin Schmid, and Michael Bowling. Low-variance and zero-variance baselines for extensive-form games. *arXiv preprint arXiv:1907.09633*, 2019.

[de Witt *et al.*, 2019] Christian Schroeder de Witt, Jakob Foerster, Gregory Farquhar, Philip Torr, Wendelin Boehmer, and Shimon Whiteson. Multi-agent common knowledge reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 9927–9939, 2019.

[Fagin *et al.*, 2003] Ronald Fagin, Yoram Moses, Joseph Y Halpern, and Moshe Y Vardi. *Reasoning about knowledge*. MIT press, 2003.

[Foerster *et al.*, 2019] Jakob Foerster, Francis Song, Edward Hughes, Neil Burch, Iain Dunning, Shimon Whiteson, Matthew Botvinick, and Michael Bowling. Bayesian action decoder for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 1942–1951. PMLR, 2019.

[Horák and Bošanský, 2019] Karel Horák and Branislav Bošanský. Solving partially observable stochastic games with public observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2029–2036, 2019.

[Johanson *et al.*, 2011] Michael Johanson, Kevin Waugh, Michael Bowling, and Martin Zinkevich. Accelerating best response calculation in large extensive games. In *IJCAI*, volume 11, pages 258–265, 2011.

[Kovařík *et al.*, 2022] Vojtěch Kovařík, Martin Schmid, Neil Burch, Michael Bowling, and Viliam Lisỳ. Rethinking formal models of partially observable multiagent decision making. *Artificial Intelligence*, 303:103645, 2022.

[Lerer *et al.*, 2020] Adam Lerer, Hengyuan Hu, Jakob N Foerster, and Noam Brown. Improving policies via search in cooperative partially observable games. In *AAAI*, pages 7187–7194, 2020.

[Li *et al.*, 2020] Hui Li, Kailiang Hu, Shaohua Zhang, Lin Wang, Jun Zhou, Yuan Qi, and Le Song. Regret minimization via novel vectorized sampling policies and exploration. *preprint*, 2020. Accessed: August 5th, 2020.

[Moravčík *et al.*, 2017] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[Schrittwieser *et al.*, 2019] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, Go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019.

[Seitz *et al.*, 2019] Dominik Seitz, Vojtěch Kovařík, Viliam Lisý, Jan Rudolf, Shuo Sun, and Karel Ha. Value functions for depth-limited solving in imperfect-information games beyond poker. *arXiv preprint arXiv:1906.06412*, 2019.

[Silver *et al.*, 2017] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354, 2017.

[Spaan *et al.*, 2008] Matthijs TJ Spaan, Frans A Oliehoek, and Nikos Vlassis. Multiagent planning under uncertainty with stochastic communication delays. In *338 Proceedings of the Eighteenth International Conference on Automated Planning and Scheduling (ICAPS 2008)*, pages 338–345, 2008.

[Šustr *et al.*, 2019] Michal Šustr, Vojtěch Kovařík, and Viliam Lisý. Monte Carlo continual resolving for online strategy computation in imperfect information games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 224–232. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[von Neumann and Morgenstern, 1953] John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton university press, 1953.