Proceedings of Machine Learning Research vol 144:1–21, 2022

Cautious Markov Games as a Framework for Human-Robot Interaction

Rohan Sinha RHNSINHA@STANFORD.EDU
Sanjay Lall LALL@STANFORD.EDU

Abstract

Safe interaction between autonomous vehicles (AVs) and humans requires decision-making strategies that intelligently reason about the effect an AV's decisions have on actors in its environment. As a result, recent years have seen much progress towards interaction-aware methods that consider the non-cooperative nature of the human-robot interactions in autonomous driving. In contrast with previous work, we propose to exploit the fact that humans and autonomous agents typically interact in environments with rules and conventions that all agents should follow, such as in traffic. To do this, we express the rules as linear temporal logical constraints on the joint state trajectory and model the multi-agent interaction as a stochastic game. We fix the likelihood of other agents making decisions that can violate the rules instead of implicitly encoding it in the agents' preference structure. This formulation results in decision-making agents that take the likelihood of others breaking the rules into account in an interpretable manner based on quantities that are straightforward to estimate. We dub this framework the *cautious Markov game* (CMG), for which we efficiently construct policies using robust dynamic programming. We also show that classic results on Nash equilibria in the 2-player zero-sum setting extend to the CMG. By exploiting the rule-based nature of the game, we can significantly reduce the conservatism of robust policies on simple illustrative examples.

Keywords: Markov Decision Processes, Formal Methods, Game Theory

1. Introduction

To reliably deploy the next generation of autonomous robotic systems in unstructured open-world environments, they have to interact safely with other agents such as humans and other robots. In many of these settings, such as when an autonomous vehicle merges into traffic on a highway on-ramp, automated systems need to negotiate with other agents in their environment to achieve an objective that is often in tension with the goals of others.

As surveyed in Schwarting et al. (2018), a rapidly growing body of research in the controls and AI communities seeks to develop methods for autonomous decision-making that reason intelligently about the influence that an autonomous agents' decisions have on the behavior of others in its environment, often by employing a variety of learning algorithms. However, recent methods ignore that interactions between humans and robots are typically highly structured: Vehicles in traffic should obey traffic rules. Therefore, in contrast with previous work, we propose taking advantage of the fact that humans and robots interact in structured environments with rules and conventions that everyone should follow. Indeed, it is our view that the interaction-aware AV decision-making task would be significantly simplified if everyone obeyed traffic rules at all times, because many traffic rules (like driving on the right-hand side of the road) exist precisely to reduce the amount of negotiation between agents that is required to safely reach a destination. Therefore, we hypothesize that estimates of the likelihood that other agents will break the rules present a highly interpretable and actionable signal to incorporate in an AV decision-making stack.

Contributions: In this work, we propose to synthesize decision making-agents that explicitly take the likelihood of others violating traffic rules into account. To do this, we propose to model the interaction rules using Linear Temporal Logical (LTL) constraints on the joint trajectories of all the agents in a stochastic game. Even though traffic rules encode statements on trajectories—did agent 1 stop at the stop-sign before crossing? Did agent 1 yield for agent 2?—and thus on the outcomes of joint decisions of all the agents, traffic rules generally make it unambiguous to humans which decisions are acceptable and which are not in a specific situation. Therefore, under the assumption that agents are unilaterally responsible for themselves satisfying traffic rules, we translate the LTL trajectory constraints into constraints on the agents' actions by segmenting each agent's action set into a subset of *prudent actions* and its *imprudent* complement using reachability analysis. We then model the likelihood of an agent breaking the rules as the probability of taking an *imprudent action*, which can result in a rule violation, and estimate this likelihood from data. We dub the resulting game the *cautious Markov game* and discuss how classic results for normal-form games translate to our setup, where the likelihood of an agent taking a rule-breaking action is fixed a priori. This allows us to solve the 2-player zero-sum stochastic game analogously to the classical work of Shapley (1953) and compute adversarially robust policies for N-player games. Finally, we show on simulated examples that our assumptions accurately model common traffic scenarios and significantly reduce the conservatism of adversarial approaches by taking traffic rules into account. As such, this work presents both a basic contribution towards intelligent, interaction-aware decision-making and an initial step towards formal synthesis methods for autonomous agents in multi-agent, rule-based environments, a problem that has received comparatively little attention in the literature Schwarting et al. (2018).

2. Related Work

Interaction-aware methods have received increasing attention from robotics researchers in recent years, resulting in a diverse body of existing work that we broadly segment into three categories, all of which do not present a principled treatment of traffic rules.

1. Ego-Conditioned Trajectory Forecasting: These methods apply contemporary deep-learning techniques to predict the joint evolution of the trajectories of all agents in a scene Ivanovic and Pavone (2019) or explicitly condition forecasts of other agents' trajectories on the motion plans of the ego vehicle Salzmann et al. (2020), thereby capturing the interaction between the agents. By conditioning the predictions on the decisions made by the ego vehicle, these methods can reason about how others will react to the decisions made by the AV. However, it is challenging to collect data with the AV behaving in many different ways in the same scenario and equip modern deep learning methods with provably correct measures of uncertainty in the predictions.

2. Intention-Aware Decision Making: These methods model the interaction-aware decision-making problem as one of estimating the intent of other agents online by considering the policies of opposing agents as functions of unknown intent parameters, reducing the problem to a single agent POMDP Bandyopadhyay et al. (2013); Sadigh et al. (2016a); Brechtel et al. (2011); Bai et al. (2015); Galceran et al. (2015). Online inference on the intentions or goals that other agents may have can be computationally prohibitive in complex interactions, so these methods typically encode primitive behaviors that depend only on a handful of parameters.

3. Game Theoretic Planning: Recent work advocates a game-theoretic approach to interaction-aware robotics, modeling opposing agents as rational with respect to a cost function to construct Nash equilibrium policies that anticipate how opponents will react Le Cleac'h et al. (2021); Spica et al. (2020); Fridovich-Keil et al. (2020); Sadigh et al. (2016b); Ye et al. (2020); Zanardi et al. (2021); Wang et al. (2019); Britzelmeier and Dreves (2020); Dreves and Gerdts (2018). However, computing Nash equilibria for general non-cooperative games is intractable Shoham and Leyton-Brown (2008). Only a handful of dynamic games, like the zero-sum (adversarial) stochastic game, have a unique solution that we can tractably compute Shapley (1953); Başar and Olsder (1998). Therefore, existing game-theoretic algorithms generally focus on practical iterative methods without strong guarantees, optimizing free-form trajectories in generic dynamic games. Many of these methods assume the objectives of other agents are known a priori (i.e., see Spica et al. (2020); Fridovich-Keil et al. (2020); Le Cleac'h et al. (2021)) or identify parametric models, like neural networks or weighted basis functions, to capture the objectives of other road users, for example, through inverse reinforcement learning algorithms Ng and Russell (2000); Sadigh et al. (2016b); Peters et al. (2021). Therefore, existing algorithms encode the tendency of an agent to follow traffic rules entirely within their objective, leading to methods that can both be fragile under uncertainty and uninterpretable while making adversarial (i.e., zero-sum) approaches much too conservative. In this work, we explicitly encode the structure in the control task by fixing the likelihood that other agents break traffic rules to reduce the amount of negotiation that is required by the agents.

Formal Methods for Decision-Making: oftentimes traffic rules specify the order in which events should occur, so we model them as linear temporal logical (LTL) Baier and Katoen (2008) constraints on the joint trajectories of the agents, in the spirit of recent work that transcribed traffic rules into LTL Rizaldi and Althoff (2015); Esterle et al. (2020). Previously, formal methods have also been applied in the controls and robotics communities to guarantee that autonomous systems satisfy complex temporal behavior Sadigh et al. (2014); Li et al. (2017); Wolff et al. (2012); Raman et al. (2015). These algorithms synthesize policies that satisfy LTL constraints for all time in various single-agent or cooperative multi-agent settings, using both exact solution methods and a variety of reinfocement learning approaches (i.e., see Baier and Katoen (2008); Wolff et al. (2012); Sadigh et al. (2014); Li et al. (2017); Sun et al. (2020); Raman et al. (2015); Leung et al. (2021); DeCastro et al. (2020)). These methods generally cannot handle the noncooperative and multi-agent nature of rule-based open-world robotics tasks, as we cannot control whether the agents surrounding an autonomous vehicle follow traffic rules or not. Therefore, we focus instead on synthesizing agents that avoid breaking the rules themselves but take the likelihood of others violating rules at any given instant into account.

3. The Cautious Stochastic Game

Now, we introduce the problem formulation for which we develop efficient solution algorithms. Consider a discrete stochastic (or Markov) game $(\mathcal{S}, T, \{\mathcal{A}_i, R_i\}_{i=1}^N, \gamma)$ as considered in Shoham and Leyton-Brown (2008), with N agents that should each follow LTL traffic rules. Specifically, let $\phi_1, \ldots, \phi_N$ be the LTL specifications that encode the traffic rule that each respective agent $i \in \{1, \ldots, N\}$ should follow, defined over atomic propositions that are functions of the state $s \in \mathcal{S}$. Unfortunately, sometimes the agents decide to break the rules, which we will simply model using a fixed probability $p_i \in [0, 1]$ for each agent $i \in \{1, \ldots, N\}$. To optimize their utility, the agents need to take the likelihood of others breaking the rule into account. We dub this formulation the *Cautious Markov Game*.

3.1. Product Game

To construct the Cautious Markov Game, we first take advantage of the defining property of the LTL formulae $\phi_1, \ldots, \phi_N$: That we can easily verify whether the state history at some time $t \in \mathbb{N}$, defined as $h^t = (s^0, \ldots, s^t)$ with $s^k \in \mathcal{S}$ as the state at time k, satisfies the rules using finite state machines (FSMs). Formally, for every rule ϕ_i, there exists a FSM $(\mathcal{Q}_i, \mathcal{S}, q_i^0, \delta_i, F_i)$ so that when the FSM is given initial condition q_i^0 and follows the transition dynamics $\delta_i : \mathcal{Q}_i \times \mathcal{S} \to \mathcal{Q}_i$, it holds that h^t satisfies ϕ_i, that is, $h^t \models \phi_i$, if and only if $q_i^{t+1} \in F_i \subseteq \mathcal{Q}_i$ Baier and Katoen (2008); De Giacomo and Vardi (2013). We therefore augment the state $s \in \mathcal{S}$ using the FSM states associated with the rules. This is a standard approach in the literature (e.g. see Sadigh et al. (2014); Wolff et al. (2012); Baier and Katoen (2008) for a detailed discussion).

Definition 1 *Let $(\mathcal{Q}_i, \mathcal{S}, q_i^0, \delta_i, F_i)$ be the FSM associated with rule ϕ_i for $i = 1, \ldots, N$. The product game associated with a stochastic game and a set of rules $\phi_1, \ldots, \phi_N$ is the tuple $(\mathcal{S}_p, T_p, \{\mathcal{A}_i, R_i\}_{i=1}^N, \gamma)$, where the product state $s_p \in \mathcal{S}_p = \mathcal{S} \times \mathcal{Q}_1, \ldots, \mathcal{Q}_N$ and the product dynamics $T_p : \mathcal{S}_p \times \mathcal{S}_p \times \mathcal{A} \to [0, 1]$ are given as*

$$T_p(s_p' | s_p, a) = \begin{cases} T(s'|s, a) & \text{if } \delta_i(q_i, s') = q_i' \quad \forall\, i \in \{1, \ldots, N\} \\ 0 & \text{else} \end{cases}. \tag{1}$$

The product game's dynamics are essentially equivalent to that of the original cautious game: They simply represent the joint evolution of the system state s, which the agents observe, and the FSM states $q_1, \ldots, q_N$, which the agents track internally. However, we can now focus our attention on memoryless policies of the product state s_p, as a trajectory history $h^t \models \phi_i$ if and only if $s_p^t = (s^t, q_1^t, \ldots, q_N^t)$ has $q_i^t \in F_i$.

3.2. Traffic Rule Desiderata

Before we fully introduce the Cautious Markov Game, we need to develop a precise notion of what it means for an individual agent to break a rule because violation of an LTL specification over state trajectories depends on the outcomes of decisions made by multiple agents. Moreover, even though arbitrary rules can make it unclear who was at fault for a rule violation from a causal modeling perspective Pearl (2009), traffic rules generally make it unambiguous how road-users should act. Therefore, we construct our approach around two basic observations:

1. Rules are discriminative: We are guided by the intuition that acceptable interaction rules should be discriminative in nature. For example, a traffic rule should be satisfied by default if all the agents decide to stay at home. As such, agents only violate the rules when they engage in behavior that was explicitly disallowed, making it possible to pinpoint the instant at which a rule was violated as the timestep after which an LTL formula was longer satisfied. This intuition means a good rule should not require some event to happen eventually, as this would imply that all the agents are breaking the rule until this event happens.

2. Rules eliminate coordination: In this work, we take the view that traffic rules generally exist to reduce the need for coordination among the agents. For example, traffic rules specify that we drive on the right-hand side of the road so that we do not have to coordinate with oncoming traffic. This perspective implies that an agent violating a rule is a result of their own decisions regardless of the behavior of others, unambiguously specifying acceptable behavior for each agent individually. Therefore, we define the event of an agent breaking traffic rules as follows:

Definition 2 *Agent $i \in \{1, \ldots, N\}$ breaks its associated* LTL *rule ϕ_i at time $t \in \mathbb{N}$ if for $k \in \{0, \ldots, t-1\}$ it holds that the trajectory history $h^k \models \phi_i$ and $h^t \not\models \phi_i$.*

From the FSM instantiation of the rules, we see that agent i breaks its associated traffic rule ϕ_i if and only if the product state $s_p \in \mathcal{S}_p$ evolves into the set

$$\mathcal{F}_i := \{(s, q_1, \ldots, q_N) \in \mathcal{S}_p : q_i \notin F_i\}, \tag{2}$$

where F_i is the set of accepting conditions of the FSM associated with ϕ_i, under definition 2.

Under our desiderata, each agent carries unilateral responsibility for breaking its associated traffic rule; if agent i runs a red light or does not yield at a roundabout, we consider this a consequence of agent i's decisions and not the other agents'. Moreover, to decide when an agent breaks a rule, we need to assume that the agent can unilaterally make decisions to satisfy the rule for all time from a nonempty set of initial conditions; nobody should be blamed for something that was out of their control from the start. We will examine common traffic scenarios in §5 that support our hypothesis—that these desiderata model the vast majority of traffic rules.

3.3. The Cautious Markov Game

To avoid breaking the rule ϕ_i, agent i needs to ensure that the state $s_p^t \notin \mathcal{F}_i$ for all time $t \geq 0$. We note that recognizing the timestep at which a rule was broken is separate from the question of when violation of the rule became inevitable: running a red light may become inevitable if a car drives too fast to be able to stop in time, but the light will not be run until the car enters the intersection.

Therefore, we identify the set $\mathcal{R}_i \subseteq \mathcal{S}_p$ from which agent i can unilaterally avoid reaching $\mathcal{F}_i$ for all future timesteps. Let Π_i be the set of all Markov policies in the product state space, i.e., functions $\pi_i : \mathcal{S}_p \to \mathcal{A}_i$, for agent i. Then, we define the set

$$\mathcal{R}_i := \{s_p \in \mathcal{S}_p : \exists \pi_i \in \Pi_i \text{ s.t. } \text{Prob}(s_p^t \in \mathcal{F}_i | s_p^0 = s_p) = 0 \; \forall \pi_{-i} \in \Pi_{-i},\; t \geq 0\}, \tag{3}$$

as the set of states from which agent i can guarantee ϕ_i is satisfied for all time and for all opponent policies. The set $\mathcal{R}_i$ therefore reflects the set of states from which agent i can unilaterally satisfy its associated traffic rule without coordinating with other agents, making precise the unilateral responsibility as noted in our desiderata.

We use $\mathcal{R}_i$ to translate the violation of agent i's traffic rule ϕ_i, a statement over realized state trajectories $h^t \in \mathcal{S}^t$, to a statement over agent i's decision making. Specifically, we define the set of *prudent*, or *good*, actions for agent i at product state $s_p \in \mathcal{S}_p$ as

$$\mathcal{G}_i(s_p) := \{a_i \in \mathcal{A}_i : T_p(s'_p | s_p, a_i, a_{-i}) = 0 \; \forall s'_p \in \mathcal{S}_p \setminus \mathcal{R}_i,\; a_{-i} \in \mathcal{A}_{-i}\}, \tag{4}$$

and we take the set of *imprudent*, or *bad*, actions as its complement $\mathcal{B}_i(s_p) := \mathcal{A}_i \setminus \mathcal{G}_i(s_p)$. For a state $s_p \in \mathcal{R}_i$, the set $\mathcal{G}_i(s_p)$ collects all the actions for which agent i remains in $\mathcal{R}_i$ with probability 1, regardless of the decisions of the other agents. Conversely, if agent i takes an *imprudent action*, there is a nonzero probability that the state evolves into $\mathcal{S}_p \setminus \mathcal{R}_i$, from which agent i can no longer guarantee that the rule will be satisfied for all time unless it coordinates with the other agents. We emphasize that taking an imprudent action does not immediately imply that agent i's rule will be broken. For example, even though an agent decided to drive too fast to yield for an opponent on a roundabout, the yield rule will not end up being broken if the opponent takes an earlier exit at the last moment. Still, under our desiderata, agent i should not take imprudent actions, as they will be at fault if a rule violation occurs. To make this intuition precise, we prove in Appendix B that agent i's rule is satisfied for all time and opponent policies $\pi_{-i} \in \Pi_{-i}$ if and only if the initial condition $s_p^0 \in \mathcal{R}_i$ and $a_i^t \in \mathcal{G}_i(s_p^t)$ for all $t \geq 0$.

Therefore, if agent i breaks rule ϕ_i, this means agent i took at least one *imprudent action*, that is, an action that can result in a future rule violation. Therefore, we simply consider the probability that an agent takes an imprudent action as a fixed quantity, perhaps estimated from data, giving rise to the *Cautious Stochastic Game* that we consider in this work.

Definition 3 *A Cautious Markov Game (CMG) is the tuple $(\mathcal{S}_p, T_p, \{\mathcal{A}_i, \mathcal{G}_i, \mathcal{B}_i, R_i, p_i\}_{i=1}^N, \gamma)$. Here $\mathcal{S}_p$ is the shared set of product states, $\mathcal{A}_i$ is the action set for agent i, $T_p : \mathcal{S}_p \times \mathcal{S}_p \times \mathcal{A} \to [0, 1]$ is the state transition probability, $\gamma \in (0, 1)$ is the discount factor, $R_i : \mathcal{S}_p \times \mathcal{A} \to \mathbb{R}$ is agent i's reward function, and*

- *$\mathcal{G}_i(s_p) \subseteq \mathcal{A}_i$ and $\mathcal{B}_i(s_p) = \mathcal{A}_i \setminus \mathcal{G}_i(s_p)$ are the sets of* prudent *and* imprudent *actions for agent i at a state $s_p \in \mathcal{S}_p$,*
- *$p_i \in [0, 1]$ is the probability that agent i takes an* imprudent *action, i.e. an action in $\mathcal{B}_i(s_p)$, whenever $\mathcal{B}_i(s_p) \neq \emptyset$ and $\mathcal{G}_i(s_p) \neq \emptyset$.*

The difference between a cautious game and a normal-form stochastic game is that the agents take imprudent actions with a fixed likelihood p_i. This formalism means that the agents need to take the likelihood of their opponents breaking the rules into account to maximize their expected cumulative reward and act *cautiously* towards their opponents, which is why we dubbed this formulation a cautious Markov game. It is a classic result for normal-form stochastic games that memoryless equilibrium policies exist Shoham and Leyton-Brown (2008). Similarly, we consider the problem of identifying Markov perfect equilibria of the cautious Markov game.

Definition 4 *Let $\Delta_i^G(s_p) = \{z_i \in \Delta(\mathcal{A}_i) : \mathrm{Prob}(a_i \in \mathcal{G}_i(s_p)) = 1 - p_i\}$ be the set of strategies for agent i in which agent i takes an imprudent action with probability p_i at a state s_p. For a Cautious Stochastic Game, a Markov perfect policy profile $\pi^\star(s_p) \in \Delta_G(s_p) := \prod_{i=1}^N \Delta_i^G(s_p)$ constitutes a Nash equilibrium (NE) if and only if*

$$\pi_i^\star(s) \in \underset{\pi_i(s_p)\in\Delta_i^G(s_p)}{\arg\max} \left\{\mathbb{E}_{\pi_i,\pi_{-i}^\star}[Q_i^{\pi^\star}(s_p, a_i, a_{-i})] =: Q_i^{\pi^\star}(s_p, \pi_i(s_p), \pi_{-i}^\star(s_p))\right\}, \quad \forall s_p \in \mathcal{S}_p, \quad (5)$$

for agents $i = 1, \ldots, N$.

4. Proposed Approach

In this section we build up our approach towards the cautious Markov game that we defined in §3. We divide our approach into three parts: First, we show how to compute $\mathcal{R}_i$ for each agent $i \in \{1, \ldots, N\}$, to define a Cautious Markov Game associated with an MG and rules $\phi_1, \ldots, \phi_N$. Then, we analyze these games in the static setting without dynamics. Finally, we use the results from the static setting to construct algorithms to efficiently solve for Nash equilibria in the 2-player zero-sum case and compute robust (maxmin) policies for a general N-agent setting.

4.1. Identifying Rule-Breaking and Rule-Following Actions

To identify the rule-breaking and rule-following action sets $\mathcal{B}_i(s_p)$ and $\mathcal{G}_i(s_p)$ necessary to formulate the Cautious Markov Game, we need to compute the safe set of states $\mathcal{R}_i$ associated with each agent $i \in \{1, \ldots, N\}$. Identifying the set $\mathcal{R}_i$ generally involves solving a *reach-avoid game*, where the ego's objective is to steer clear of $\mathcal{F}_i$ and the opponent's disturbance maximizes the likelihood of entering $\mathcal{F}_i$. For multi-agent MDPs, this can be done by taking advantage of classic results for zero-sum stochastic games Shapley (1953). Similarly, for continuous time systems, sets like $\mathcal{R}_i$ defined in (3) are often computed using Hamilton-Jacobi reachability analysis Bansal et al. (2017).

We make a simple observation to reduce computational cost: Since there are a finite number of decisions that the other agents can make, we can perform any robust (i.e. for all opponent actions) reachability computation for agent i by *uniformly randomizing* over the decisions made by the opposing agents.

Definition 5 *We write the transition dynamics associated with randomizing the actions of all agents except i as*

$$T_i(s_p'|s_p, a_i) := \frac{1}{|\mathcal{A}_{-i}|} \sum_{a_{-i}\in\mathcal{A}_{-i}} T_p(s_p'|s_p, a_i, a_{-i}). \quad (6)$$

Using the randomized transition function, we then solve a single agent reachability problem.

Theorem 6 *Define the MDP $(\mathcal{S}_p, T_i, \mathcal{A}_i, R_i^{\text{reach}}, \gamma)$ as a reachability problem for each agent $i = 1, \ldots, N$, with*

$$R_i^{\text{reach}}(s_p, a_i) = \begin{cases} -1 & \text{if } s_p \in \mathcal{F}_i \\ 0 & \text{else} \end{cases} \quad (7)$$

and $\gamma \in (0, 1)$. Let π_i^{reach}, V_i^{reach} be the optimal policy and value function associated with this MDP. Then, it holds that

$$\mathcal{R}_i = \{s_p \in \mathcal{S}_p : V_i^{\text{reach}}(s_p) = 0\}. \quad (8)$$

For a proof of Theorem 6, see Appendix C. Therefore, we can compute $\mathcal{R}_i$, and by extension $\mathcal{G}_i(s_p)$ and $\mathcal{B}_i(s_p)$, efficiently using single-agent value iteration.

4.2. Games with Priors Over Actions

To build out the machinery to handle stochastic games with rule-breaking likelihoods, we first discuss how definition 4 extends the Nash equilibrium concept for static games without dynamics. In our setting, we know the likelihood of agent i taking an action in $\mathcal{B}_i$ is p_i a priori, so we say that these games have a *prior over actions*.

Clearly, the set of mixed profiles that satisfy the prior over actions, $\pi_i \in \Delta_i^G$ and, by extension, $\Delta_G := \prod_{i=1}^N \Delta_i^G$ are convex. Although definition 4 appears similar to the regular definition of a mixed Nash equilibrium, the constraint set $\Delta_G \subsetneq \Delta := \prod_{i=1}^N \Delta(\mathcal{A}_i)$ is different. The prior over actions indicates some partial knowledge of the behavior of other agents, such as an estimate of the likelihood that a road user will act to violate a traffic rule. There need not exist any mixed Nash equilibria to the game that happen to satisfy the prior over actions, whereas it is trivial to show that an equilibrium according to definition 4 has to exist for a game with priors over actions by repeating Nash's classic fixed-point proof, since Δ_G is a convex subset of Δ Nash (1950). Therefore, specifying games with priors over actions allows us to align the result of game theoretic analysis with partial observations on an opponent's behavior without modifying the incentive structures of the agents. Moreover, by selecting $p_i = 0$ and $\mathcal{G}_i = \mathcal{A}_i$, it should be readily apparent that games with priors over actions generalize normal-form games. Therefore, general games with priors over actions inherit at least the computational hardness of general-sum normal-form games (specifically, PPAD-completeness Roughgarden (2016)), so we consider efficiently solving these games in a general setting intractable.

Because identifying equilibria for general-sum N-player games is a computationally intractable problem, we instead focus on computing worst-case, or robust, strategies and payoffs for each agent. To compute robust payoffs for agent i, we assume that all the other agents centrally coordinate their decisions to frustrate agent i, while still satisfying the prior over actions.

Definition 7 *For a game with priors over actions, let* $\Delta_{-i}^R = \{z_{-i} \in \Delta(\mathcal{A}_{-i}) : \mathrm{Prob}(a_j \in \mathcal{B}_j) = p_j \quad \forall j \neq i\}$*. Then, the robust payoff* $R_i^\star$ *and robust strategy* $z_i^\star \in \Delta_i^G$ *for agent* $i \in \{1, \ldots, N\}$ *are the solution and optimizer of the (maxmin) problem*

$$R_i^\star = \max_{z_i \in \Delta_i^G} \min_{z_{-i} \in \Delta_{-i}^R} R_i(z_i, z_{-i}). \tag{9}$$

In Appendix E we show that we can solve (9) using linear programming (LP) in two separate ways: one uses using LP duality, and one takes advantage of an interpretation involving p_i-biased coin tosses. Because Δ_{-i}^G parametrizes only strategy profiles where the opponents behave independently of each other and Δ_{-i}^R does not, we get a lower-bound certificate of performance when agent i plays strategy $z_i^\star$. That is, $R_i^\star \leq R_i(z_i^\star, z_{-i})$ for all $z_{-i} \in \Delta_{-i}^G \subseteq \Delta_{-i}^R$.

It is a classic result in game theory that robust strategies correspond to the unique Nash equilibrium for 2-player zero-sum games Dasgupta et al. (2008); Shoham and Leyton-Brown (2008). This classic result also applies to games with priors over actions, we include a proof in Appendix D.

4.3. Cautious Stochastic Games

Now that we analyzed the static, or zero-shot, setting with no dynamics as games with priors over actions, we return to the cautious stochastic games we wish to solve. In definition 4, the stage games played at each state $s_p \in \mathcal{S}_p$ constitute games with priors over actions. As a result, computing NEs for CMGs using a value-iteration-like algorithm is computationally hard. Moreover, it is well known that, even if we assume we could compute equilibria to the stage games, a value iteration algorithm need not converge to a NE Greenwald and Hall (2003). Instead, we compute policies for our autonomous agent, the ego agent, using a *robust value-iteration* (VI) procedure. In a robust

VI scheme, we assume the other agents coordinate together against the ego. Therefore we solve for the robust policy and value function that satisfy equation (9) at each state at each iteration. We summarize this procedure in algorithm 1. In comparison with the classical Value Iteration algorithm, in robust VI, we solve an LP at each state for each iteration. We can therefore view algorithm 1 as a generalization of the classic algorithm proposed for 2-player zero-sum Markov Games in Shapley (1953).

Lemma 8 *Algorithm 1 converges to a unique fixed point $V_i^\star$ for each agent $i = 1, \dots, N$.*

For a proof of lemma 16, see Appendix F. Moreover, since algorithm 1 is a value iteration procedure and we can efficiently compute the robust problem (9) using LP, it is straightforward to see that we can efficiently compute an approximate fixed point. In addition, we show that in the 2-player zero-sum case, algorithm 1 identifies the unique NE of the CMG. This is an unsurprising generalization of the classic result due to Shapley (1953). For a proof, see Appendix F. Moreover, we emphasize that the robust value function $V_i^\star$ and its associated implicit policy $\pi_i^\star$ produce a certificate of robustness for agent i, that is, that agent i's expected utility will be at least $V_i^\star$ under the implicit robust policy.

Theorem 9 *Let $V_i^\star$ be the output of algorithm 1 for agent i, with implicit maxmin policy $\pi_i^\star$ such that $\pi_i^\star(s_p) \in \Delta_i^G(s_p)$ at each $s_p \in \mathcal{S}_p$. Let V_i be the value function for agent i under $\pi_i^\star$ and any opponent policies π_{-i} such that $\pi_{-i}(s_p) \in \Delta_{-i}^G(s_p)$ at each $s_p \in \mathcal{S}_p$. Then it holds for any $s_p \in \mathcal{S}_p$ that $V_i^\star(s_p) \leq V_i(s_p)$.*

For a proof, see Appendix F. We note that for general Markov games without rules, the robust policy computed for an ego robot is often much too conservative to find practical use. Especially in autonomous driving, a collision penalty for the ego agent will then result in a robust policy that assumes other road users will try to collide with the ego agent. However, by specifying the game's rules, the CMG constrains the other agents only to take actions that violate the rules with a fixed probability. Hence, if the rules are well-designed, much of the ambiguity of the other agents' behavior is eliminated a priori, reducing the conservatism of the robust policy without applying inverse RL approaches and searching for general-sum Nash equilibria.

```
Algorithm 1: Cautious Markov Game Robust Value Iteration
Given: CMG ($\mathcal{S}_p, \gamma, T_p, \{R_i, \mathcal{G}_i, \mathcal{B}_i, p_i\}_{i=1}^N$), ego agent $i$;
$V_i^0(s_p) \leftarrow 0 \quad \forall s_p \in \mathcal{S}_p$ and $k \leftarrow 1$;
while not converged do
    for $s_p \in \mathcal{S}_p, a \in \mathcal{A}$ do
        $\mathcal{Q}_i^k(s_p, a) \leftarrow R_i(s_p, a) + \gamma \sum_{s_p' \in \mathcal{S}_p} T_p(s_p'|s_p, a) V_i^{k-1}(s_p')$;
    end
    for $s_p' \in \mathcal{S}_p$ do
        $V_i^k(s_p) \leftarrow$ solution to (9) using $\mathcal{Q}_i^k(s_p, \cdot)$;
    end
    $k \leftarrow k + 1$;
end
return $\mathcal{Q}_i^\star(s_p, a), V_i^\star(s_p)$
```

```
Algorithm 2: Learning in Cautious Markov Games
Given: MG ($\mathcal{S}, T, R_i, \gamma, \{\mathcal{A}_i\}_{i=1}^N$), ego agent $i$, trajectory dataset $\mathcal{D} = \{\{s^j, a^j\}_{j=1}^{T_k}\}_{k=1}^K$. ;
Construct product dynamics as per definition 1.;
$\mathcal{R}_j \leftarrow$ solution to reachability problem (12) for $j = 1, \dots, N$. ;
$\mathcal{G}_j(s_p), \mathcal{B}_j(s_p) \leftarrow$ equation (4) $\forall s_p \in \mathcal{S}_p$, $j = 1, \dots, N$.
$\hat{p}_j(s_p) \leftarrow$ estimate of $p_j$ from $\mathcal{D}$ for $j \neq i$. ;
Compute $\pi_i^\star(s_p)$ using algorithm 1 and $\{\hat{p}_j\}_{j \neq i}$.
```

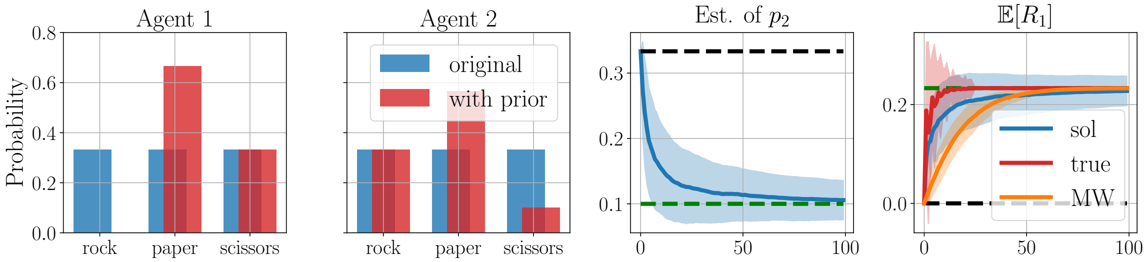

Figure 1: Left two plots: NE policies for RPS for the original zero-sum game, and with a prior over actions on agent 2. Center right: Estimate of $\mathrm{Prob}(a_2 = \text{scissors})$ vs. the no. of games played. Right: Expected reward of agent 1: sol indicates $R^\star$ from (9) under the current estimate, true compares the current policy to π_2, and MW compares a policy computed using the MW algorithm to π_2, all with 2σ bars.

Moreover, the rule-violation probabilities p_i for all non-ego agents are straightforward to estimate from a dataset of K trajectories $\mathcal{D} = \{\{s^j, a^j\}_{j=1}^{T_k}\}_{k=1}^{K}$, by counting the events when $a_i \in \mathcal{G}_i(s_p)$ in the product state space. Moreover, we emphasize that our results apply not only when p_i is fixed, but also when it is a function of the product state, i.e., when $p_i : \mathcal{S}_p \to [0, 1]$, so that p_i can be a local property of the environment. This parallels the standard result that $\mathcal{A}_i$ can also be a function of the state. We summarize this approach in algorithm 2.

5. Simulations

Rock-Paper-Scissors (RPS): First off, we illustrate the utility of a prior over actions with a simple static example. Consider a game of rock-paper-scissors between two agents, the ego and the opponent. Each agent receives a payoff of 1 if it wins, 0 if the outcome is a stalemate, and -1 if it loses, a classic zero-sum game. As is shown in Fig. 1, the *unique* NE for rock-paper-scissors is for both players to uniformly randomize their strategies, yielding an expected payoff of 0 for each agent Dasgupta et al. (2008). Suppose now that we play many RPS games in sequence and only observe our reward and whether the human opponent played scissors or not. We might find that they do not randomize their play exactly according to the zero-sum NE. For example, as shown in Fig. 1, we might observe over time that $\mathrm{Prob}(a_2 = \text{scissors}) = \frac{1}{10}$.

Rather than identifying a payoff function that *explains* our opponent's behavior, resulting in a non zero-sum game requiring non-convex optimization to compute a NE, we can *align* the structure of the game with observations of our opponent's behavior by specifying a game with a prior over actions. As shown in Fig. 1, specifying the prior allows us to quickly exploit partial knowledge of our actual opponent's behavior, as we quickly learn not to play rock, resulting in an expected payoff of .23 and a 41% chance of winning. Moreover, Fig. 1 also shows that online estimation of the prior over actions allows us to learn optimal behavior faster than if we apply the multiplicative weights (MW) algorithm (which even requires observing a_2 at each iteration) Roughgarden (2016).

Grid World: Next, we illustrate the cautious Markov game and our solution algorithms on a grid world example with two agents, the ego and the opponent (see Fig. 2). Our grid world example represents a simplified depiction of a stop-signed intersection. Both agents can only move in straight lines along the grid cells; the ego moves vertically and the opponent horizontally. The middle cell at the $(0, 0)$ coordinate, where the directions of motion of both agents intersect, represents the intersection that both agents need to cross. The Markovian dynamics of the agents are independent, each of which is described by its position and velocity $s_i = (x_i, v_i)$, and acceleration command $a_i \in (-1, 0, 1)$. We construct the transition dynamics as follows: Each agent moves its position to

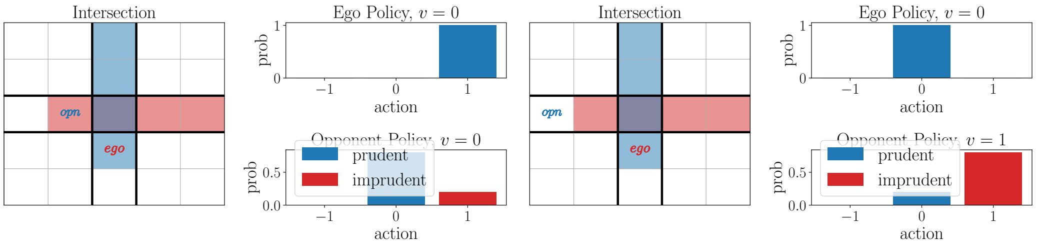

Figure 2: Behavior of the cautious grid-world agents. Left: $p_2 = \frac{1}{5}$. Right: $p_2 = \frac{4}{5}$. The shaded regions in the intersection indicate the states in $\mathcal{S}_p \setminus \mathcal{R}_i$ for the ego (blue) and opponent (red) reachable from $s_1^0 = s_2^0 = (-2, 0)$. The bar charts show the agents' policies.

	$p_2 = 0$	$p_2 = \frac{1}{5}$	$p_2 = \frac{4}{5}$	$p_2 = 1$
Cautious	5.3	3.2	.16	0.0
Optimist	5.3	3.2	-.95	-2.0
Pessimist	1.2	0.8	.11	0.0

Table 1: Realised ego utility $V_1(s_p^0)$ against an opponent that takes imprudent actions with probability p_2.

$x_i + v_i$ with $v_i \in (-1, 0, 1)$, with a fixed transition probability $p_t = \frac{1}{2}$ and stays in the same place with probability $1 - p_t$. The velocity v_i' at the next timestep is equal to the acceleration command a_i with probability 1. The agents receive a positive reward for reaching $x_i = 2$ before their opponent and are penalized for collisions. Both agents are expected to respect the first-in-first-out (FIFO) traffic principle, necessitating an (ϕ Before ϕ') operator, which we construct from the base LTL operators as **SB** (read as *strictly before*) in Appendix A. We take the atomic propositions A_i, B_i, C_i as indicators on whether agents $1, 2$ occupy the grid cell right before the intersection (that is, $(0, -1)$ and $(-1, 0)$), the intersection cell $(0, 0)$, or have crossed the intersection (any grid coordinate greater than 0) respectively and state the FIFO rules

$$\phi_i = (A_j \textbf{ SB } A_i) \rightarrow (C_j \textbf{ SB } B_i),$$

for $(i, j) \in \{(1, 2), (2, 1)\}$. Our ego agent should never break the rules, so we set $p_1 = 0$. Fig. 2 shows the qualitative behavior of the agents when the ego arrives at the stop sign first: the ego decides to cross the road if p_2 is low, but lets the opponent pass when they are likely to take imprudent actions and cross out of turn. Moreover, for $s_1^0 = (-1, 0)$ and $s_2^0 = (-2, 1)$, Table 1 shows that the cautious agent 1) is less conservative than a pessimistic agent that assumes $p_2 = 1$, and 2) performs better than an optimistic agent that assumes $p_2 = 0$, thereby assuming agent 2 always takes prudent actions.

6. Conclusion and Future Work

In this work, we developed a precise notion of what it means to for an agent to break a traffic rule in a stochastic game. First, we used this definition to convert rule-breaking, a statement over trajectories, to one over the agents' decisions. Then, given an estimate of the likelihood that agents take imprudent actions, we discussed how facts about normal-form games generally translate to the setting in which we know this prior over actions. Our simulations showed that accounting for the rule-based nature of the interaction through the prudent and imprudent actions can significantly reduce the conservatism of robust (max-min) policies on a simple example. This showcases that 1) traffic rules reduce the need for inter-agent negotiation 2) estimates of when opponents make imprudent decisions are an actionable signal in an AV stack. Therefore, we focus future work towards conducting a wide range of case studies on data that closely models the real world, for example, in the CARLA simulator Dosovitskiy et al. (2017).

References

H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee. Intention-aware online POMDP planning for autonomous driving in a crowd. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 454–460, May 2015. doi: 10.1109/ICRA.2015.7139219. ISSN: 1050-4729.

Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. The MIT Press, Cambridge, Mass, 2008. ISBN 978-0-262-02649-9. OCLC: ocn171152628.

Tirthankar Bandyopadhyay, Kok Sung Won, Emilio Frazzoli, David Hsu, Wee Sun Lee, and Daniela Rus. Intention-Aware Motion Planning. In *Algorithmic Foundations of Robotics X*, volume 86, pages 475–491. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-36278-1 978-3-642-36279-8. doi: 10.1007/978-3-642-36279-8_29. URL http://link.springer.com/10.1007/978-3-642-36279-8_29. Series Title: Springer Tracts in Advanced Robotics.

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J. Tomlin. Hamilton-jacobi reachability: A brief overview and recent advances, 2017.

Tamer Başar and Geert Jan Olsder. *6. Nash and Saddle-Point Equilibria of Infinite Dynamic Games*, pages 265–363. 1998. doi: 10.1137/1.9781611971132.ch6. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611971132.ch6.

Sebastian Brechtel, Tobias Gindele, and Rüdiger Dillmann. Probabilistic MDP-behavior planning for cars. In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1537–1542, October 2011. doi: 10.1109/ITSC.2011.6082928. ISSN: 2153-0017.

Andreas Britzelmeier and Axel Dreves. A decomposition algorithm for Nash equilibria in intersection management. *Optimization*, 0(0):1–38, June 2020. ISSN 0233-1934. doi: 10.1080/02331934.2020.1786088. URL https://doi.org/10.1080/02331934.2020.1786088. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02331934.2020.1786088.

Sanjoy Dasgupta, Christos Papadimitriou, and Umesh Vazirani. *Algorithms*. McGraw-Hill, 2008. ISBN 978-0-07-352340.

Giuseppe De Giacomo and Moshe Y. Vardi. Linear Temporal Logic and Linear Dynamic Logic on Finite Traces. In *IJCAI '13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 854–860. Association for Computing Machinery, 2013. URL https://scholarship.rice.edu/handle/1911/78495. Accepted: 2014-11-21T22:06:43Z.

Jonathan DeCastro, Karen Leung, Nikos Arechiga, and Marco Pavone. Interpretable Policies from Formally-Specified Temporal Properties. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7, Rhodes, Greece, September 2020. IEEE. ISBN 978-1-72814-149-7. doi: 10.1109/ITSC45102.2020.9294442. URL https://ieeexplore.ieee.org/document/9294442/.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

Axel Dreves and Matthias Gerdts. A generalized Nash equilibrium approach for optimal control problems of autonomous cars. *Optimal Control Applications and Methods*, 39(1):326–342, 2018. ISSN 1099-1514. doi: 10.1002/oca.2348. URL http://onlinelibrary.wiley.com/doi/abs/10.1002/oca.2348. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/oca.2348.

Klemens Esterle, Luis Gressenbuch, and Alois Knoll. Formalizing Traffic Rules for Machine Interpretability. *2020 IEEE 3rd Connected and Automated Vehicles Symposium (CAVS)*, pages 1–7, November 2020. doi: 10.1109/CAVS51000.2020.9334599. URL http://arxiv.org/abs/2007.00330. arXiv: 2007.00330.

David Fridovich-Keil, Ellis Ratner, Lasse Peters, Anca D. Dragan, and Claire J. Tomlin. Efficient iterative linear-quadratic approximations for nonlinear multi-player general-sum differential games. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1475–1481, 2020. doi: 10.1109/ICRA40945.2020.9197129.

Enric Galceran, Alexander Cunningham, Ryan Eustice, and Edwin Olson. Multipolicy Decision-Making for Autonomous Driving via Changepoint-based Behavior Prediction. In *Robotics: Science and Systems XI*. Robotics: Science and Systems Foundation, July 2015. ISBN 978-0-9923747-1-6. doi: 10.15607/RSS.2015.XI.043. URL http://www.roboticsproceedings.org/rss11/p43.pdf.

Amy Greenwald and Keith Hall. Correlated Q-Learning. page 8, 2003.

Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Simon Le Cleac'h, Mac Schwager, and Zachary Manchester. ALGAMES: A Fast Augmented Lagrangian Solver for Constrained Dynamic Games. *Autonomous Robots*, 2021.

Karen Leung, Nikos Aréchiga, and Marco Pavone. Back-propagation through Signal Temporal Logic Specifications: Infusing Logical Structure into Gradient-Based Methods. *arXiv:2008.00097 [cs, eess]*, January 2021. URL http://arxiv.org/abs/2008.00097. arXiv: 2008.00097.

Xiao Li, Cristian-Ioan Vasile, and Calin Belta. Reinforcement learning with temporal logic rewards. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3834–3839, Vancouver, BC, September 2017. IEEE. ISBN 978-1-5386-2682-5. doi: 10.1109/IROS.2017.8206234. URL http://ieeexplore.ieee.org/document/8206234/.

John F. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950. ISSN 0027-8424. doi: 10.1073/pnas.36.1.48.

Andrew Y. Ng and Stuart Russell. Algorithms for Inverse Reinforcement Learning. In *in Proc. 17th International Conf. on Machine Learning*, pages 663–670. Morgan Kaufmann, 2000.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Lasse Peters, David Fridovich-Keil, Vicenç Rubies-Royo, Claire Tomlin, and Cyrill Stachniss. Inferring objectives in continuous dynamic games from noise-corrupted partial state observations. 07 2021. doi: 10.15607/RSS.2021.XVII.030.

Vasumathi Raman, Alexandre Donzé, Dorsa Sadigh, Richard M. Murray, and Sanjit A. Seshia. Reactive synthesis from signal temporal logic specifications. In *Proceedings of the 18th International Conference on Hybrid Systems: Computation and Control*, HSCC '15, pages 239–248, New York, NY, USA, April 2015. Association for Computing Machinery. ISBN 978-1-4503-3433-4. doi: 10.1145/2728606.2728628. URL https://doi.org/10.1145/2728606.2728628.

Albert Rizaldi and Matthias Althoff. Formalising Traffic Rules for Accountability of Autonomous Vehicles. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 1658–1665, Gran Canaria, Spain, September 2015. IEEE. ISBN 978-1-4673-6596-3. doi: 10.1109/ITSC.2015.269. URL http://ieeexplore.ieee.org/document/7313361/.

Tim Roughgarden. *Twenty Lectures on Algorithmic Game Theory*. Cambridge University Press, Cambridge, 2016. ISBN 978-1-107-17266-1. doi: 10.1017/CBO9781316779309. URL https://www.cambridge.org/core/books/twenty-lectures-on-algorithmic-game-theory/A9D9427C8F43E7DAEF8C702755B6D72B.

D. Sadigh, E. S. Kim, S. Coogan, S. S. Sastry, and S. A. Seshia. A learning based approach to control synthesis of Markov decision processes for linear temporal logic specifications. In *53rd IEEE Conference on Decision and Control*, pages 1091–1096, December 2014. doi: 10.1109/CDC.2014.7039527. ISSN: 0191-2216.

D. Sadigh, S. S. Sastry, S. A. Seshia, and A. Dragan. Information gathering actions over human internal state. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 66–73, October 2016a. doi: 10.1109/IROS.2016.7759036. ISSN: 2153-0866.

Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. Planning for Autonomous Cars that Leverage Effects on Human Actions. page 9, 2016b.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision – ECCV 2020*, pages 683–700, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58523-5.

Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and Decision-Making for Autonomous Vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1(1): 187–210, 2018. doi: 10.1146/annurev-control-060117-105157. URL https://doi.org/10.1146/annurev-control-060117-105157.

L. S. Shapley. Stochastic Games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100, October 1953. Publisher: National Academy of Sciences Section: Mathematics.

Yoav Shoham and Kevin Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511811654.

Riccardo Spica, Davide Falanga, Eric Cristofalo, Eduardo Montijano, Davide Scaramuzza, and Mac Schwager. A Real-Time Game Theoretic Planner for Autonomous Two-Player Drone Racing. *IEEE Transactions on Robotics*, 36(5), 2020.

C. Sun, X. Li, and C. Belta. Automata Guided Semi-Decentralized Multi-Agent Reinforcement Learning. In *2020 American Control Conference (ACC)*, pages 3900–3905, July 2020. doi: 10.23919/ACC45564.2020.9147704. ISSN: 2378-5861.

Mingyu Wang, Zijian Wang, John Talbot, J. Christian Gerdes, and Mac Schwager. Game Theoretic Planning for Self-Driving Cars in Competitive Scenarios. In *Robotics: Science and Systems XV*. Robotics: Science and Systems Foundation, June 2019. ISBN 978-0-9923747-5-4. doi: 10.15607/RSS.2019.XV.048. URL http://www.roboticsproceedings.org/rss15/p48.pdf.

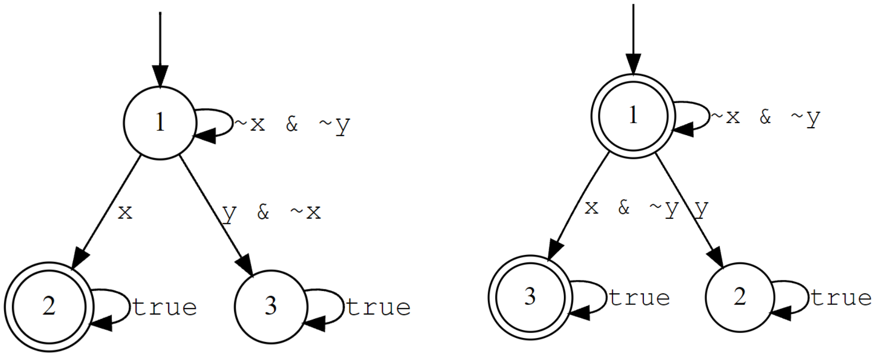

Figure 3: Left: Finite state machine associated with the loose before operator. Right: FSM associated with the strict before operator. Initial state is indicated with the arrow from the top, accepting conditions are represented with doubly circled nodes.

E. M. Wolff, U. Topcu, and R. M. Murray. Robust control of uncertain Markov Decision Processes with temporal logic specifications. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 3372–3379, December 2012. doi: 10.1109/CDC.2012.6426174. ISSN: 0743-1546.

G. Ye, Q. Lin, T.-H. Juang, and H. Liu. Collision-free Navigation of Human-centered Robots via Markov Games. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11338–11344, May 2020. doi: 10.1109/ICRA40945.2020.9196810. ISSN: 2577-087X.

Alessandro Zanardi, Enrico Mion, Mattia Bruschetta, Saverio Bolognani, Andrea Censi, and Emilio Frazzoli. Urban Driving Games With Lexicographic Preferences and Socially Efficient Nash Equilibria. *IEEE Robotics and Automation Letters*, 6(3):4978–4985, July 2021. ISSN 2377-3766. doi: 10.1109/LRA.2021.3068657. Conference Name: IEEE Robotics and Automation Letters.

Appendix A. Defining the "Before" Operator in LTL.

Many traffic rules and conventions specify certain events should happen in the right order, if they are to occur at all. An ability to express a requirement that one event should occur *before* another is indispensable to properly encode real-world traffic rules like the FIFO principle at a stop-signed intersection. We show such an operator can directly be constructed from the base operators $\mathbf{U}$, $\mathbf{X}$ in LTL syntax Baier and Katoen (2008). However, some nuance exists in the definition of a "before" operator: For example, should an operator that requires "x before y" allow x and y to occur at the same instant or not? We define both a strict and a loose "before" operator to express either requirement.

Definition 10 *For two* LTL *formulae x and y, we define the* loose before *operator as* $x\mathbf{LB}y := (\neg y)\mathbf{U}x$. *If an input trace $h \models x\mathbf{LB}y$, we say x happens loosely before y.*

Definition 11 *For two* LTL *formulae x and y, we define the* strict before *operator as* $x\mathbf{SB}y := \neg(y\mathbf{LB}x) = \neg((\neg x)\mathbf{U}y)$. *If an input trace $h \models x\mathbf{SB}y$, we say x happens strictly before y.*

We illustrate the behavior of the loose and strong before operators using their corresponding finite state machines, shown in Fig. 3. Besides the fact that both operators encode the basic idea that x should occur before y, it is apparent that the *loose before* operator

1. allows x and y to be satisfied for the first time simultaneously,

2. requires x to be satisfied at some point in time.

In contrast, the *strict before* operator

1. does not require x or y to be satisfied at some point in time,

2. requires x to be satisfied strictly before y is satisfied for the first time.

Since the *strict before* operator is accepting by default and disallows events happening simultaneously, it presents a useful tool in modelling real-world traffic rules that satisfy our desiderata.

Appendix B. Prudent and Imprudent actions

Lemma 12 *If $s_p \in \mathcal{R}_i$, then there exists an action $a_i \in \mathcal{A}_i$ so that $T_p(s'_p|s_p, a_i, a_{-i}) = 0$ for all $s'_p \in \mathcal{S}_p \setminus \mathcal{R}_i$ and $a_{-i} \in \mathcal{A}_{-i}$. In other words, there exists an action a_i so that $\operatorname{Prob}(s'_p \in \mathcal{R}_i) = 1$.*

Proof Since s_p is in $\mathcal{R}_i$, there exists a policy π_i so that $\mathcal{F}_i$ is avoided for all time for all $\pi_{-i} \in \Pi_{-i}$ almost surely. Therefore, if agent i starts from s_p, takes the action $a_i = \pi_i(s_p)$, and transitions into s'_p, then $s'_p \notin \mathcal{F}_i$ and agent i can avoid $\mathcal{F}_i$ from s'_p by continuing to act according to π_i. Therefore, there exists a policy so that $\mathcal{F}_i$ is avoided from s'_p, implying that $s'_p \in \mathcal{R}_i$. ■

Theorem 13 *Define the set of* prudent*, or* good*, actions at each state $s_p \in \mathcal{S}_p$ for each agent $i = 1, \ldots, N$ as the set*

$$\mathcal{G}_i(s) := \{a \in \mathcal{A}_i : T_p(s'_p|s_p, a_i, a_{-i}) = 0 \; \forall s'_p \in \mathcal{S}_p \setminus \mathcal{R}_i,\ a_{-i} \in \mathcal{A}_{-i}\}, \tag{10}$$

and let the set of imprudent*, or* bad *actions be $\mathcal{B}_i(s_p) := \mathcal{A}_i \setminus \mathcal{G}_i(s_p)$. Then, for a policy $\pi_i \in \Pi_i$, it holds that $\operatorname{Prob}(s_p^t \in \mathcal{F}_i|s_p^0 = s_p) = 0$ for all $t \geq 0$ and $\pi_{-i} \in \Pi_{-i}$ if and only if the initial condition $s_p^0 \in \mathcal{R}_i$ and $\pi_i(s_p^t) \in \mathcal{G}_i(s_p^t)$ for all $t \geq 0$.*

Proof We first show the "only if" direction. If $s_p^t \in \mathcal{R}_i$, then lemma 12 implies that $\mathcal{G}_i(s_p^t)$ is not empty. By definition, if $s_p^t \in \mathcal{R}_i$, taking $\pi_i(s_p^t) \in \mathcal{G}_i$ then implies that $s_p^{t+1} \in \mathcal{R}_i$. Moreover, $\mathcal{F}_i \cap \mathcal{R}_i = \emptyset$, so if $s_p^0 \in \mathcal{R}_i$, then the unilateral policy π_i for agent i guarantees $\mathcal{F}_i$ is avoided by induction.

Now we show the other direction. Assume that under $\pi_i \in \Pi_i$, the state never enters $\mathcal{F}_i$ from some initial condition for any opponent policy $\pi_{-i} \in \Pi_{-i}$. We proceed by contradiction: Suppose that there exists some $t \geq 0$ and some realizable trajectory history $h_p^t \in \mathcal{S}_p^t$ under π_i from s_p^0 for which $a_i^t = \pi_i(s_p^t) \notin \mathcal{G}_i(s_p^t)$. Then, by the definition of $\mathcal{G}_i(s_p^t)$, it holds that $\operatorname{Prob}(s_p^{t+1} \notin \mathcal{R}_i) > 0$ for some $a_{-i} \in \mathcal{A}_{-i}$. By the definition of $\mathcal{R}_i$, there are no policies in Π_i that can guarantee that $\mathcal{F}_i$ is avoided for an initial condition in $\mathcal{S}_p \setminus \mathcal{R}_i$ for an arbitrary $\pi_{-i} \in \Pi_{-i}$. Therefore, there then exists some $k \geq 0$ and $\pi_{-i} \in \Pi_{-i}$ so that $\operatorname{Prob}(s_p^{t+k+1} \in \mathcal{F}_i|s_p^0) > 0$ under the policy π_i. This contradicts our premise, proving the theorem. ■

Appendix C. Proof of Theorem 6

Theorem 14 *Define the MDP $(\mathcal{S}_p, T_i, \mathcal{A}_i, R_i^{\text{reach}}, \gamma)$ as a reachability problem for each agent $i = 1, \dots, N$, with*

$$R_i^{\text{reach}}(s_p, a_i) = \begin{cases} -1 & \text{if } s_p \in \mathcal{F}_i \\ 0 & \text{else} \end{cases} \tag{11}$$

and $\gamma \in (0, 1)$. Let π_i^{reach}, V_i^{reach} be the optimal policy and value function associated with this MDP. Then, it holds that

$$\mathcal{R}_i = \{s_p \in \mathcal{S}_p : V_i^{\text{reach}}(s_p) = 0\}. \tag{12}$$

Proof First, notice that for $(s_p', s_p, a_i) \in \mathcal{S}_p \times \mathcal{S}_p \times \mathcal{A}_i$ it holds that

$$T_i(s_p'|s_p, a_i) = 0 \iff T(s_p'|s_p, a_i, a_{-i}) = 0 \quad \forall a_{-i} \in \mathcal{A}_{-i}. \tag{13}$$

Now, we prove the "only if" direction. Let $\hat{\mathcal{R}}_i = \{s_p \in \mathcal{S}_p : V_i^{\text{reach}}(s_p) = 0\}$. Since

$$V_i^{\text{reach}}(s_p) = R_i^{\text{reach}}(s_p, \pi_i^{\text{reach}}(s_p)) + \gamma \sum_{s_p' \in \mathcal{S}_p} T_i(s_p'|s_p, \pi_i^{\text{reach}}(s_p) V_i^{\text{reach}}(s_p')$$

and $R_i^{\text{reach}}(s_p, a_i) \leq 0$, it follows that $V_i^{\text{reach}}(s_p) = 0$ if and only if $V_i^{\text{reach}}(s_p') = 0$ for all $s_p' \in \mathcal{S}_p$ for which $T_i(s_p'|s_p, \pi_i^{\text{reach}}(s_p)) > 0$. Therefore, for a state $s_p \in \hat{\mathcal{R}}_i$, it holds that $T_i(s_p'|s_p, \pi_i^{\text{reach}}(s_p)) = 0$ for all $s_p' \in \mathcal{S}_p \setminus \hat{\mathcal{R}}_i$. Using (13), it follows that for $s_p \in \hat{\mathcal{R}}_i$, it holds that $T_p(s_p'|s_p, \pi^{\text{reach}}(s_p), a_{-i}) = 0$ for all $s_p' \in \mathcal{S}_p \setminus \hat{\mathcal{R}}_i$ and $a_{-i} \in \mathcal{A}_{-i}$. Since $\hat{\mathcal{R}}_i \cap \mathcal{F}_i = \emptyset$, we therefore have that if $s_p \in \hat{\mathcal{R}}_i$, there exists a policy such that $\text{Prob}_{T_p}(s_p^t \in \mathcal{F}_i | s_p^0 = s_p) = 0$ for all $t \geq 0$ and $\pi_{-i} \in \Pi_{-i}$ by induction. Therefore $\hat{\mathcal{R}}_i \subseteq \mathcal{R}_i$.

For the other direction, assume that $s_p \in \mathcal{R}_i$. Therefore, there exists a policy π_i so that $\text{Prob}_{T_p}(s_p^t \in \mathcal{F}_i | s_p^0 = s_p) = 0$ for all $t \geq 0$ and $\pi_{-i} \in \Pi_{-i}$. By theorem 13, this implies that for $s_p \in \mathcal{R}_i$, it holds that $\pi_i(s_p) \in \mathcal{G}_i(s_p)$. Therefore, we have that $T_p(s_p'|s_p, \pi_i(s_p), a_{-i}) = 0$ for all $s_p' \in \mathcal{S}_p \setminus \mathcal{R}_i$ and $a_{-i} \in \mathcal{A}_{-i}$. Equation (13) then implies that $T_i(s_p'|s_p, \pi_i(s_p)) = 0$ for all $s_p' \in \mathcal{S}_p \setminus \mathcal{R}_i$. Therefore, if the initial condition $s_p \in \mathcal{R}_i$, the policy π_i guarantees that $s_p^t \in \mathcal{R}_i$ for all $t \geq 0$ under the randomized dynamics $T_i(s_p'|s_p, \pi_i(s_p))$ by induction. This implies that the value function under π_i and the randomized dynamics $T_i(\cdot)$ satisfies $V_i^{\pi}(s_p) = 0$. Since we assumed $V_i^{\text{reach}}(s_p) \in [\frac{-1}{1-\gamma}, 0]$ is the optimal value function of the reachability problem, it holds that $V_i^{\text{reach}}(s_p) \geq V_i^{\pi}(s_p) = 0$. Therefore, we have that $\mathcal{R}_i \subseteq \hat{\mathcal{R}}_i$, proving the result. ■

Appendix D. 2-Player Zero-Sum Games

Theorem 15 *Suppose we have a 2-player zero-sum game with priors over actions. That is, suppose $R_1(a) = -R_2(a) =: R(a)$ for all $a \in \mathcal{A}$. Then, there exists a unique Nash equilibrium $R^\star = R_1^\star = -R_2^\star$ that is attained when both players play max-min strategies $z_1^\star$ and $z_2^\star$, defined as*

$$z_i^\star \in \arg\max_{z_i \in \Delta_i^G} \min_{z_j \in \Delta_j^G} R_i(z_i, z_j), \tag{14}$$

for $(i, j) \in \{(1, 2), (2, 1)\}$.

Proof Suppose $\hat{z}_1 \in \Delta_1^G$ and $\hat{z}_2 \in \Delta_2^G$ form a Nash equilibrium. Since the game is zero sum, $R_1(\hat{z}) = -R_2(\hat{z}) =: \hat{R}$. Applying the definition of Nash equilibrium to agent 1, it holds that

$$\hat{R} = \max_{z_1 \in \Delta_1^G} R(z_1, \hat{z}_2) \geq \min_{z_2 \in \Delta_2^G} \max_{z_1 \in \Delta_1^G} R(z_1, z_2). \tag{15}$$

Similarly, using the NE definition and the fact that $R_2(a) = -R_1(a) = -R(a)$, we have for agent 2 that

$$\hat{R} = \min_{z_2 \in \Delta_2^G} R(\hat{z_1}, z_2) \leq \max_{z_1 \in \Delta_1^G} \min_{z_2 \in \Delta_2^G} R(z_1, z_2). \tag{16}$$

If we define the matrix $\mathbf{R} \in \mathbb{R}^{|\mathcal{A}_1| \times |\mathcal{A}_2|}$, with entries $[\mathbf{R}]_{ij} = R(a_1^i, a_2^j)$ for $a_1^i \in \mathcal{A}_1$ and $a_2^j \in \mathcal{A}_2$, then it holds that $R(z_1, z_2) = z_1^\intercal \mathbf{R} z_2$ for $(z_1, z_2) \in \Delta_G$. Therefore, by von Neumman's minimax theorem and convexity of the compact sets Δ_1^G and Δ_2^G, defining $R^\star$ as

$$R^\star := \min_{z_2 \in \Delta_2^G} \max_{z_1 \in \Delta_1^G} z_1^\intercal \mathbf{R} z_2 = \max_{z_1 \in \Delta_2^G} \min_{z_2 \in \Delta_1^G} z_1^\intercal \mathbf{R} z_2, \tag{17}$$

implies that $R^\star \leq \hat{R} \leq R^\star$. Hence, the game has a unique equilibrium attained when the agents play maxmin policies.

■

Appendix E. Adversarially Robust Strategies for N-player games

In this section, we show how to compute the robust value and policy associated with a game with priors over actions defined in definition 7 in two ways. Both methods show that the maxmin problem is a Linear Program.

E.1. Base Interpretation

From the definition of Δ_{-i}^R, we can write the robust problem (9) as

$$\begin{aligned} \max_{z_i \in \Delta_i^G} \min_{z_{-i}} \; & z_i^\intercal \mathbf{R}_i z_{-i} \\ \text{s.t.} \quad & z_{-i} \in \Delta(\mathcal{A}_{-i}), \\ \text{Prob}_{z_{-i}}(a_j \in \mathcal{B}_j) &= p_j \quad \forall j \neq i, \end{aligned}$$

where the matrix $\mathbf{R}_i \in \mathbb{R}^{|\mathcal{A}_i| \times |\mathcal{A}_{-i}|}$ has entries $[\mathbf{R}_i]_{kj} = R_i(a_i^k, a_{-i}^j)$ for $a_i^k \in \mathcal{A}_i$ and $a_{-i}^j \in \mathcal{A}_{-i}$. We will write the inner min as a max by taking the dual. First, define $b^\intercal = [1, p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_N]$. Also, let the matrix

$$\mathbf{F}_i = \begin{bmatrix} f_0 & f_1 & \dots & f_{i-1} & f_{i+1} & \dots & f_N \end{bmatrix} \in \mathbb{R}^{|\mathcal{A}_{-i}| \times N},$$

where $f_0 = \mathbb{1}_{|\mathcal{A}_{-i}|}$ is the one vector and the k'th entry of the vector $[f_j]_k = 1$ if $a_{-i}^k = (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_N)$ has $a_j \in \mathcal{B}_j$. Then we can rewrite the inner min as

$$\begin{aligned} \min_{z_{-i}} \; & \underbrace{z_i^\intercal \mathbf{R}_i}_{=:c^\intercal} z_{-i} \\ \text{s.t.} \quad & z_{-i} \geq 0 \\ & \mathbf{F}_i^\intercal z_{-i} = b. \end{aligned}$$

The lagrangian of this LP is

$$\begin{aligned}\mathcal{L}(z_{-i},\lambda,\nu) &= c^\intercal z_{-i} + \lambda^\intercal(-z_{-i}) + \nu^\intercal(\mathbf{F}_i^\intercal z_{-i} - b)\\ &= (c + \mathbf{F}_i\nu - \lambda)^\intercal z_{-i} - \nu^\intercal b,\end{aligned}$$

from which it follows that the Lagrangian dual function is

$$g(\lambda,\nu) := \inf_{z_{-i}} \mathcal{L}(z_{-i},\lambda,\nu) = \begin{cases} -\nu^\intercal b & \text{if } (c + \mathbf{F}_i\nu - \lambda) = 0 \\ -\infty & \text{else} \end{cases}.$$

Hence, the dual program $\max_{\lambda\geq 0,\ \nu} g(\lambda,\nu)$ can be rewritten as the LP

$$\begin{aligned}&\max_{\lambda,\ \nu} - \nu^\intercal b\\ &\text{s.t.} \quad \lambda \geq 0\\ &c + \mathbf{F}_i\nu = \lambda,\end{aligned}$$

which is equivalent to

$$\begin{aligned}&\max_{\nu} -\nu^\intercal b\\ &\text{s.t.} \quad \mathbf{R}_i^\intercal z_i + \mathbf{F}_i\nu \geq 0.\end{aligned}$$

Substituting $v = -\nu$, we get

$$\begin{aligned}&\max_{v} v^\intercal b\\ &\text{s.t.} \quad \mathbf{R}_i^\intercal z_i - \mathbf{F}_i v \geq 0.\end{aligned}$$

returning to the original problem, it follows that

$$\begin{aligned}R_i^\star = &\max_{z_i\in\Delta_i^G} \max_{v} v^\intercal b\\ &\text{s.t.} \quad \mathbf{R}_i^\intercal z_i - \mathbf{F}_i v \geq 0\end{aligned} \tag{18}$$

which is a basic LP with decision variables z_i and ν. In addition, note that the optimal profile $z_i^\star$ that maximizes (18) is also a maximimzer for the robust problem (9).

E.2. Random Agent Interpretation

The second method relies on the interpretation that a game-with priors over actions corresponds to a game where each agent first flips a coin to decide whether to play a prudent or imprudent action. To compute the robust value $R_i^\star$, we consider a coordinated agent for every outcome scenario of the coin tosses. To explicitly write this out, let the set $\mathcal{X}_j = \{(1-p_j,\mathcal{G}_j),(p_j,\mathcal{B}_j)\}$ collect the tuples representing the prudent and imprudent scenarios for each agent. In addition, let $\mathcal{X}_{-i} = \prod_{j\neq i}\mathcal{X}_j$ collect every opponent scenario for agent i. Then, for every scenario $x_{-i} \in \mathcal{X}_{-i}$, define the associated opponent action set as $\mathcal{U}_{-i}(x_{-i}) = \prod_{(w_j,\mathcal{U}_j)\in x_j \forall j\neq i}\mathcal{U}_j$. This action set occurs with likelihood $w_{x_{-i}} = \prod_{(w_j,\mathcal{U}_j)\in x_j \forall j\neq i} w_j$. By representing the coordinated opponent strategy using distributions conditioned over each scenario, it follows that

$$\Delta_{-i}^R = \big\{\{w_{x_{-i}} z_{x_{-i}}\}_{x_{-i}\in\mathcal{X}_{-i}} : z_{x_{-i}} \in \Delta(\mathcal{U}_{x_{-i}}) \quad \forall x_{-i} \in \mathcal{X}_{-i}\big\}.$$

It then follows that the robust problem

$$\max_{z_i \in \Delta_i^G} \min_{z_{-i} \in \Delta_{-i}^R} R_i(z_i, z_{-i})$$

is equivalent to

$$\begin{aligned} \max_{z_i} \min_{z_{x_{-i}}} & \sum_{x_{-i} \in \mathcal{X}_{-i}} w_{x_{-i}} R_i(z_i, z_{x_{-i}}) \\ \text{s.t.} \quad & z_i \in \Delta_i^G \\ & z_{x_{-i}} \in \Delta(\mathcal{U}_{-i}(x_{-i})) \quad \forall x_{-i} \in \mathcal{X}_{-i}. \end{aligned}$$

Now, we define the matrices $\mathbf{R}_i^{g,x_{-i}} \in \mathbb{R}^{|\mathcal{G}_i| \times |\mathcal{U}_{-i}(x_{-i})|}$ and $\mathbf{R}_i^{b,x_{-i}} \in \mathbb{R}^{|\mathcal{B}_i| \times |\mathcal{U}_{-i}(x_{-i})|}$ with k, l'th entries as $R_i(a_i^k, a_{-i}^l)$ for $a_i^k \in \mathcal{G}_i$, $a_{-i}^l \in \mathcal{U}_i(x_{-i})$ and $a_i^k \in \mathcal{B}_i$, $a_{-i}^l \in \mathcal{U}_i(x_{-i})$ respectively. Then, applying the epigraph trick, it follows that we can write the robust problem as

$$\begin{aligned} & \max_{x_i, y_i, t_{x_{-i}}} \sum_{x_{-i} \in \mathcal{X}_{-i}} w_{x_{-i}} t_{x_{-i}} \\ \text{s.t.} \quad & x_i \in \Delta(\mathcal{G}_i), \quad y_i \in \Delta(\mathcal{B}_i), \quad t_{x_{-i}} \in \mathbb{R} \quad \forall x_{-i} \in \mathcal{X}_{-i}, \\ & t_{x_{-i}} \leq (1-p_i) \mathbf{R}_i^{g,x_{-i}\intercal} x_i + p_i \mathbf{R}_i^{b,x_{-i}\intercal} y_i, \quad \forall x_{-i} \in \mathcal{X}_{-i}. \end{aligned} \tag{19}$$

Appendix F. Robust Value Iteration for Cautious Games

Lemma 16 *Algorithm 1 converges to a unique fixed point $V_i^\star$ for each agent $i = 1, \ldots, N$.*

Proof The proof presented is essentially an analogous result to that of the original 2-player zero-sum game due to Shapley (1953). Moreover, it can also be shown using robust dynamic programming theory as presented in Wolff et al. (2012). For any function $V : \mathcal{S}_p \to \mathbb{R}$, define $\|V\|_\infty = \max_{s_p \in \mathcal{S}_p} |V(s_p)|$ and let $\text{CB}[V_i](s_p)$ denote the result at state s_p of applying one iteration of the value iteration algorithm 1 to V_i for agent $i \in \{1, \ldots, N\}$. Then, for any two $V_i, \hat{V}_i$ and $s_p \in \mathcal{S}_p$, it holds that

$$\begin{aligned} \text{CB}[V_i](s_p) &= \max_{\pi_i(s_p) \in \Delta_i^G(s_p)} \min_{\pi_{-i}(s_p) \in \Delta_{-i}^R(s_p)} Q_i(s_p, \pi_i(s_p), \pi_{-i}(s_p))) \\ &= \max_{\pi_i(s_p) \in \Delta_i^G(s_p)} \min_{\pi_{-i}(s_p) \in \Delta_{-i}^R(s_p)} \mathbb{E}_{a \sim (\pi_i(s_p), \pi_{-i}(s_p))} \big[R_i(s_p, a) + \gamma \sum_{s_p \in \mathcal{S}_p} T_p(s_p' | s_p, a) V_i(s_p') \big] \\ &= \max_{\pi_i(s_p) \in \Delta_i^G(s_p)} \min_{\pi_{-i}(s_p) \in \Delta_{-i}^R(s_p)} \mathbb{E}_{a \sim (\pi_i(s_p), \pi_{-i}(s_p))} \big[R_i(s_p, a) + \gamma \mathbb{E}_{s_p'}[V_i(s_p') | a] \big] \\ &= \max_{\pi_i(s_p) \in \Delta_i^G(s_p)} \min_{\pi_{-i}(s_p) \in \Delta_{-i}^R(s_p)} \mathbb{E}_{a \sim (\pi_i(s_p), \pi_{-i}(s_p))} \big[R_i(s_p, a) + \gamma \mathbb{E}_{s_p'}[\hat{V}(s_p') | a] + \gamma \mathbb{E}_{s_p'}[V_i(s_p') - \hat{V}_i(s_p') | a] \big] \\ &\leq \text{CB}[\hat{V}_i](s_p) + \max_{\pi_i(s_p) \in \Delta_i^G(s_p)} \max_{\pi_{-i}(s_p) \in \Delta_{-i}^R(s_p)} \gamma \mathbb{E}_{s_p'}[V_i(s_p') - \hat{V}_i(s_p')] \\ &\leq \text{CB}[\hat{V}_i](s_p) + \gamma \|V_i - \hat{V}_i\|_\infty. \end{aligned}$$

By applying a symmetric argument to $\text{CB}[\hat{V}_i](s_p)$, we get that $|\text{CB}[V_i](s_p) - \text{CB}[\hat{V}_i](s_p)| \leq \gamma \|V_i - \hat{V}_i\|_\infty$. Therefore, we have that $\text{CB}[\cdot]$ is a contraction operator; i.e., that

$$\|\text{CB}[V_i] - \text{CB}[\hat{V}_i]\|_\infty \leq \gamma \|V_i - \hat{V}_i\|_\infty,$$

since $\gamma \in (0, 1)$. By the contractive fixed point theorem Wolff et al. (2012), contractivity of $\mathrm{CB}[\cdot]$ implies that algorithm 1 converges to a unique fixed point $V_i^\star$ for which

$$V_i^\star(s_p) = \max_{z_i \in \Delta_i^G(s_p)} \min_{z_{-i} \in \Delta_{-i}^R(s_p)} Q_i^\star(s_p, z_i, z_{-i}), \quad \forall s_p \in \mathcal{S}_p.$$

■

Lemma 17 *For a 2-player zero-sum Cautious Stochastic game, a policy profile $\pi^\star$ such that $\pi^\star(s_p) \in \Delta_G(s_p)$ for all $s_p \in \mathcal{S}_p$, associated with value functions $V_1^\star(s_p)$ and $V_2^\star(s_p)$, constitutes a Nash equilibrium if and only if both players play robust (maxmin) strategies. That is, if and only if*

$$\pi_i^\star(s_p) \in \operatorname*{arg\,max}_{\pi_i(s_p) \in \Delta_i^G(s_p)} \min_{\pi_j(s_p) \in \Delta_j^G(s_p)} Q_i^\star(s_p, \pi_i(s_p), \pi_j(s_p)) \quad \forall s_p \in \mathcal{S}_p,$$

for $(i, j) \in \{(1, 2), (2, 1)\}$. Moreover, for each state $s_p \in \mathcal{S}_p$, it holds that $V_1^\star(s_p) = -V_2^\star(s_p)$.

Proof Since the game is zero sum, it follows that for any policy profile π with $\pi(s_p) \in \Delta_G(s_p)$ for all $s_p \in \mathcal{S}_p$, it holds that $V_1^\pi(s_p) = -V_2^\pi(s_p)$ for every $s_p \in \mathcal{S}_p$. Therefore, it follows that the Q-functions for each agent associated with any such policy π satisfy $Q_1^\pi(s_p, a) = -Q_2^\pi(s_p, a)$ for all $s_p \in \mathcal{S}_p$ and $a \in \mathcal{A}$. By definition 4, a policy profile $\pi^\star$ is a Nash equilibrium if and only if $\pi^\star(s_p)$ constitutes a Nash equilibrium for the subgame at each state $s_p \in \mathcal{S}_p$ with payoffs $Q_i^\star(s_p, a)$ for each agent $i = 1, 2$. Since the game is zero-sum, we have that $Q_1^\star(s, a) = -Q_2^\star(s, a)$, so these subgames are zero-sum 2-player games with priors over actions. Theorem 15 therefore gives us that a Markov policy profile for the 2-player zero-sum CMG constitutes a Nash equilibrium if and only if both agents act according to maxmin strategies, and that $V_1^\star(s_p) = -V_2^\star(s_p)$ for each $s_p \in \mathcal{S}_p$. ■

Theorem 18 *For a 2-player Cautious Stochastic Game, Algorithm 1 converges to the* unique *Nash equilibrium value function $V^\star(s_p) = V_1^\star(s_p) = -V_2^\star(s_p)$ of the game, attained when both agents play their robust (maxmin) implicit policies. That is, when*

$$\pi_i^\star(s_p) \in \operatorname*{arg\,max}_{\pi_i(s_p) \in \Delta_i^G(s_p)} \min_{\pi_j(s_p) \in \Delta_j^G(s_p)} Q_i^\star(s_p, \pi_i(s_p), \pi_j(s_p)) \quad \forall s_p \in \mathcal{S}_p,$$

for $(i, j) \in \{(1, 2), (2, 1)\}$.

Proof Since the game has 2 players, by Lemma 16, we have that Algorithm 1 converges to a unique fixed point $V_i^\star$ satisfying

$$V_i^\star(s_p) = \max_{z_i \in \Delta_i^G(s_p)} \min_{z_j \in \Delta_j^G(s_p)} Q_i^\star(s_p, z_i, z_{-i}), \quad \forall s_p \in \mathcal{S}_p, \tag{20}$$

for $(i, j) \in \{(1, 2), (2, 1)\}$. Moreover, since the game is zero-sum, Lemma 17 then yields the result. ■

Theorem 19 *Let $V_i^\star$ be the output of algorithm 1 for agent i, with implicit maxmin policy $\pi_i^\star$, i.e., a policy satisfying*

$$\pi_i^\star(s_p) \in \operatorname*{arg\,max}_{\pi_i(s_p) \in \Delta_i^G(s_p)} \min_{\pi_{-i}(s_p) \in \Delta_{-i}^R(s_p)} Q_i^\star(s_p, \pi_i(s_p), \pi_{-i}(s_p)) \quad \forall s_p \in \mathcal{S}_p.$$

Let V_i be the value function for agent i under $\pi_i^\star$ and any opponent policies π_{-i} such that $\pi_{-i}(s_p) \in \Delta_{-i}^G(s_p)$ at each $s_p \in \mathcal{S}_p$. Then it holds for any $s_p \in \mathcal{S}_p$ that $V_i^\star(s_p) \leq V_i(s_p)$.

Proof Since $\pi_i^\star$ is the implicit maxmin policy associated with $V_i^\star$, lemma 16 gives us that

$$V_i^\star(s_p) = \min_{\pi_{-i}(s_p) \in \Delta_{-i}^R(s_p)} Q_i^\star(s_p, \pi_i^\star(s_p), \pi_{-i}(s_p)) \quad \forall s_p \in \mathcal{S}_p. \tag{21}$$

Next, let π_{-i} be any coordinated opponent policy so that $\pi_{-i}(s_p) \in \Delta_{-i}^R(s_p)$ for each $s_p \in \mathcal{S}_p$, and let V_i denote the cumulative expected discounted reward for agent i under the policies $\pi_i^\star$ and π_{-i}. By Bellman's principle of optimality, equation (21) then gives us that $V_i^\star(s_p) \leq V_i(s_p)$ for all $s_p \in \mathcal{S}_p$. The result then follows from the fact that $\Delta_{-i}^G(s_p) \subseteq \Delta_{-i}^R(s_p)$ for all $s_p \in \mathcal{S}_p$. ■