GENERALIZATION GAMES IN REINFORCEMENT LEARNING

Manfred Diaz, Charlie Gauthier, Glen Berseth & Liam Paull
Mila - Québec AI Institute, Canada
Université de Montréal, Canada
{diazcabm,charlie.gauthier,glen.bertseth,paulll}@mila.quebec

ABSTRACT

In reinforcement learning (RL), the term generalization has either denoted introducing function approximation to reduce the intractability of large state and action spaces problems or designated RL agents' ability to transfer learned experiences to one or more evaluation tasks. Recently, many subfields have emerged to understand how distributions of training tasks affect an RL agent's performance in unseen environments. While the field is extensive and ever-growing, recent research has underlined that variability among the different approaches is not as significant. We leverage this intuition to demonstrate how current methods for generalization in RL are specializations of a general framework. We obtain the fundamental aspects of this formulation by rebuilding a Markov Decision Process (MDP) from the ground up by resurfacing the game-theoretic framework of games against nature. The two-player game that arises from considering nature as a complete player in this formulation explains how existing methods rely on learned and randomized dynamics and initial state distributions. We develop this result further by drawing inspiration from mechanism design theory to introduce the role of a principal as a third player that can modify the payoff functions of the decision-making agent and nature. The games induced by playing against the principal extend our framework to explain how learned and randomized reward functions induce generalization in RL agents. The main contribution of our work is the complete description of the Generalization Games for Reinforcement Learning, a multiagent, multiplayer, game-theoretic formal approach to study generalization methods in RL. We offer a preliminary ablation experiment of the different components of the framework. We demonstrate that a more simplified composition of the objectives that we introduce for each player leads to comparable, and in some cases superior, zero-shot generalization compared to state-of-the-art methods, all while requiring almost two orders of magnitude fewer samples.

1 INTRODUCTION

In reinforcement learning (RL), the term generalization has denoted the practice of introducing function approximation to reduce the intractability of problems with large state and action spaces (Sutton, 1996; Pack Kaelbing et al., 1996) or to designate RL agents' capacity to *transfer* learned experiences to one or more evaluation tasks (Whiteson et al., 2011). Recently, the fields of curriculum learning Portelas et al. (2020), open-ended learning (Wang et al., 2019), continual learning (Khetarpal et al., 2020), meta-learning (Duan et al., 2016), simulation-to-reality transfer (Sadeghi & Levine, 2016), or unsupervised environment design (Dennis et al., 2020) have emerged to understand how distributions of training tasks affect an RL agent's performance in unseen environments. While the field is extensive and ever-growing, recent surveys have pointed out that the amount of variability among the different approaches is not as significant (Portelas et al., 2020; Kirk et al., 2021). For instance, Portelas et al. (2020) managed to classify more than 30 approaches along three axes of variation: the intended use of the algorithm, aspects of the problem the algorithm controls, and the optimization objective it proposes. While the first axis is subjective and depends on the intended application of the algorithm, variations on the degree of control over the transition dynamics, initial

state distributions, and reward functions in Markov Decision Process (MDP) problems indicate the existence of a shared underlying structure of the problem.

We leverage this intuition to develop our core hypothesis that **current methods for generalization in RL are specializations of a general framework**. To uncover the fundamental aspects of this problem, we rebuild from the ground up the MDP framework (Puterman, 2005) by resurfacing and expanding *games against nature* (Milnor, 1951; Papadimitriou, 1985), a game-theoretic approach to decision-making under uncertainty. This formulation acknowledges the existence of a *nature* player that may either be a single entity or the composition of multiple ones. Those entities may possess agency and the capacity to learn and adapt and while doing so, can reshape the problem landscape of a decision-making agent (Leibo et al., 2019; Baker et al., 2020). We show that, under some assumptions, an MDP problem is a particular case of a repeated game a decision-making agent plays against a selfless nature player with unknown payoffs. At this juncture, we diverge from the traditional literature to make nature's player a *first-class citizen* within a *complete game against nature*. In this two-player game, both players aim to maximize their respective non-constant payoff (or reward) functions. The introduction of nature's payoff function prompts, under the same assumptions as before, an MDP for nature's player when it repeatedly interacts with the decision-making agent. We argue that this dual formulation explains how existing algorithms rely on *learned* or *randomized* variations of transition dynamics and initial state distributions of MDP problems to induce generalization in RL agents.

Moreover, drawing inspiration from the field of mechanism design (Vickrey, 1961), we consider the emergence of complete games against nature that assume transformations on the payoff functions of both nature and the decision-making agent. To account for these variations, we introduce the figure of the principal (Myerson, 1983), a third player that can manipulate both nature and the decision-making agent payoff structures. We first consider the role played by mechanisms where the principal is an entity external to the design process. In this context, we argue that the extensive collection of RL problems (Bellemare et al., 2012; Brockman et al., 2016) are, under the equivalences we establish, mechanism-induced repeated games against nature. Consequently, we reason that *problem designers* recreate the role of the principal by predefining the payoff structure of the decision-making agent. Then, to form a complete game against nature from these existing problems, it suffices to choose the appropriate payoff functions for nature's player. Inspired by the types of evolutionary pressures nature poses over living agents and in line with existing methods, we formulate cooperative and competitive complete agent-nature games.

Furthermore, inspired by adaptive approaches to mechanism design (Conitzer & Sandholm, 2002; Vorobeychik et al., 2007; Baumann et al., 2018), we describe *games against the principal*, a three-player game between the decision-making agent, nature, and the principal. In these games, the principal's objective is to discover the optimal outcome function from a restricted set that maximizes either player's total payoff. We propose to set the payoff structure of the principal such that it adopts a dictatorial role (Sandholm, 2003) and sides with either the decision-making agent or nature's player. In combination with the cooperative and competitive games induced by manually-design mechanism over RL problems, the optimization of the dictatorial principal objective explains how existing algorithms rely on learned or randomized reward functions to induce generalization in RL agents. Our proposal of a general framework for generalization emerges from considering the consequences rendered over a RL agent by the non-stationarity induced by nature and the principal players learning to optimize their respective payoff functions. In our presentation of the framework of **Generalization Games for Reinforcement Learning**, we concentrate on the three RL problems structured as repeated non-differentiable games against the principal, where each agent learns and execute a mixed and parametric strategy. The core contribution of this work is to demonstrate how this multiagent, multiplayer, game-theoretic framework provides a formal framework to study existing or propose novel methods leading to in RL. Notably, we show that the axes of control and optimization in the taxonomy introduced by Portelas et al. (2020) innately emerge from our formulation.

We leverage the experimental setting proposed in the context of unsupervised environment design (Dennis et al., 2020). Furthermore, we include an exploratory ablation experiment that explores the effects that coalitions with a dictatorial principal have on the final zero-shot generalization of an RL agent under adversarial and cooperative objectives. Through these experiments, we show how the more simplified composition of the objectives in our ablation study leads to comparable and,

in some cases, superior zero-shot generalization than existing state-of-the-art methods (Jiang et al., 2021) at almost two orders of magnitude fewer samples.

2 Games Against Nature and Reinforcement Learning

In this section, we resurface and expand an alternative formulation of the problem of decision-making under uncertainty, *games against nature* (Milnor, 1951; Papadimitriou, 1985; Biswas, 1997; LaValle, 2006). We describe how the MDP framework (Puterman, 2005) is a particular case of a repeated game against nature under the assumptions of perfect information and sequential nature-agent interactions. Likewise, we show that from a game-theoretic perspective, a *repeated game against nature* is an *incomplete game* where nature does not have a known payoff (or reward) function. In consequence, we introduce *complete games against nature* to consider that nature's player may also optimize for a non-constant payoff. We show that each *complete game against nature* then also induces a dual MDP problem for nature's player.

2.1 Games Against Nature

The game against nature framework presents a game-theoretic approach to decision theory and decision-making problems where a selfish and rational agent, the decision-making agent (the *agent*), aims to maximize its payoff against a disinterested opponent (*nature*) that plays at random. We derive a more formal definition of a game against nature from the works of Milnor (1951), Papadimitriou (1985), Biswas (1997), and LaValle (2006) as follows:

Definition 1 ***(Game Against Nature)*** *A game against nature is an asymmetric two-player game $\mathcal{G} = \langle \mathcal{S}, \mathcal{A}, \ell_\pi, \ell_\rho \rangle$ between a player π (the agent) and player ρ (nature) with action spaces $a \in \mathcal{A}$ and $s \in \mathcal{S}$. In this game, the player π maximizes its payoff function $\ell_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ against a* nature *player ρ that plays at random with disinterested and constant payoff $\ell_\rho : \mathcal{S} \times \mathcal{A} \to c$.*

The asymmetry in this game arises from the different action spaces of the two players. Furthermore, each player outcome is determined by their strategies $\pi \in \Delta(\mathcal{A})$ and $\rho \in \Delta(\mathcal{S})$ defined in the space of *discrete* probability distributions $\Delta(\cdot)$ over their respective action spaces $\mathcal{A}$ and $\mathcal{S}$. The traditional formulation of *games against nature* identifies several *principles of choice* that prescribe how the decision-making agent should select its strategy under uncertainty over nature's strategy (Milnor, 1951; Biswas, 1997; LaValle, 2006). From the different principles, the principles of *insufficient reason* Milnor (1951), *minimax regret* (Savage, 1951) and *expected utility* (Von Neumann & Morgenstern, 1944; Savage, 1954) are the more widespread. In our discussion, we focus on the principle of *expected utility* as it is most commonly used in the modern literature in decision-making . Under this principle, the decision-making agent π should select actions $a \in \mathcal{A}$ that maximize its expected payoff $\mathcal{L}_\pi = \mathbb{E}_{\substack{a\sim\pi \\ s\sim\rho}} \left[\ell_\pi(s, a)\right]$ under nature's mixed strategy ρ.

However, seldom if ever, games are played once. In many instances, players repeatedly interact on the same *base or stage* game, a situation modelled by *repeated games*. Similarly, a decision-making agent and nature may continually interact, leading to a repetition of the base game $\mathcal{G}$. We formally define a *repeated game against nature* as follows:

Definition 2 ***(Repeated Game Against Nature)*** *A* ***repeated game against nature*** *$\mathcal{G}^t$ is a iterated version of a base or stage game against nature $\mathcal{G}$, with $t \geq 0$ interactions between an agent π and the nature player ρ. In this game, the history of interactions $\tau = \{s_{t-1}, a_{t-1}, \ldots, a_0, s_0\} \in (\mathcal{S} \times \mathcal{A})^t$ conditions the players' payoffs functions $\ell^t_\pi : (\mathcal{S} \times \mathcal{A})^t \to \mathbb{R}$ and $\ell^t_\rho : (\mathcal{S} \times \mathcal{A})^t \to c$, respectively.*

Under the expected utility principle, the decision-making agent expected outcome $\mathcal{L}_\pi$ should be defined for *repeated games against nature* such that:

$$\mathcal{L}^t_\pi = \mathbb{E}_{p(\tau|\pi,\rho)} \left[\ell^t_\pi(\tau)\right] \tag{1}$$

where $p(\tau|\pi, \rho)$ denotes the probability distribution over feasible t-long interaction histories $\tau \in (\mathcal{S} \times \mathcal{A})^t$ conditioned also by the nature player mixed strategy ρ. There are several sensitive choices for payoff functions $\ell^t(\tau)$ for repeated players interactions in finite and infinite repeated games including average, limit of means, and discounted formulations (Abreu, 1988). We present a more detailed discussion of these functions in Appendix B.1.

2.2 MDP AND REPEATED GAMES AGAINST NATURE

We are interested in connecting the *games against nature* formalism with the MDP framework employed in RL literature. To do so, we extend the definition of *repeated game against nature* in Def. 2 to consider perfect information and sequential nature-agent interactions. First, repeated agent-nature interactions enable considering the amount of information available to each player's decision-making. We are interested in perfect information games where players can observe their past actions and those of their opponents to introduce *repeated perfect information games against nature* $\mathcal{G}^t$ where both the agent and nature strategies are conditioned on past interactions. Formally,

Definition 3 ***(Repeated Perfect Information Game Against Nature)*** *A repeated perfect information game against nature $\mathcal{G}^t$ is a repeated game against nature where each player strategy $\pi : (\mathcal{S} \times \mathcal{A})^t \to \Delta(\mathcal{A})$ and $\rho : (\mathcal{S} \times \mathcal{A})^t \to \Delta(\mathcal{S})$ is conditioned on the history $\tau \in (\mathcal{S} \times \mathcal{A})^t$ such that, at interaction time t, their actions are drawn from their mixed strategies $a_t \sim \pi(\cdot|\tau)$ and $s_t \sim \rho(\cdot|\tau)$, respectively.*

In general, *repeated games against nature* $\mathcal{G}^t$ model a game between the decision-making agent and nature where both players choose their action simultaneously. Simultaneous repeated games against nature with perfect information is a suitable realization to study *real-time RL* problems (Ramstedt & Pal, 2019). However, we focus on the more widespread practical assumption that in sequential decision-making problems *nature moves first*. For these cases, we extend *repeated perfect information games against nature* to consider sequential agent-nature interactions where nature's player starts the game, and the decision-making agent acts upon observing nature's previous action. More formally,

Definition 4 *A sequential perfect information game against nature is a perfect information game against nature where nature plays first with action $s_0 \sim \rho$ drawn from initial mixed strategy $\rho_0 \in \Delta(\mathcal{S})$. Thereafter, nature and the agent plays with strategies $\rho : (\mathcal{S} \times \mathcal{A})^t \to \Delta(\mathcal{S})$ and $\pi : \mathcal{S} \times (\mathcal{S} \times \mathcal{A})^t \to \Delta(\mathcal{A})$ defined such that, at interaction time t, the agent actions are drawn from the mixed strategy $a_t \sim \pi(\cdot|s_t, \tau)$ according to nature's last action $s_t \sim \rho(\cdot|\tau)$ and the history τ of past interactions.*

The assumptions of perfect information and sequential interaction allow us to present Markov Decision Processes MDP as a particular repeated games against nature. Note that, an MDP problem $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \rho, r, \rho_0 \rangle$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition dynamics $\rho : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$, reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and initial state distribution $\rho_0 : \mathcal{P}(\mathcal{S})$ is a sequential game a decision-making agent π plays against nature's strategy ρ. If ρ_0 denotes nature's player unconditioned strategy at interaction time $t = 0$ (nature's first move) the decision-making agent plays with strategy or policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$. Under *Markovian* assumptions (Puterman, 2005), nature's actions s_t represent sufficient statistics of past nature-agent interactions τ. Thus, the agent actions are drawn from its mixed strategy $a_t \sim \pi(\cdot|s_t)$ conditioned on nature's last action s_t. Furthermore, nature's player actions for interaction times $t \geq 1$ are drawn from a mixed strategy $s_t \sim \rho(s_t|s_{t-1}, a_{t-1})$ (the transition dynamics) and assumed independent of the rest of the interaction history τ given the last step actions s_{t-1} and a_{t-1}. Likewise, under this formulation, the reward function $r(s, a)$ is equivalent to one-shot payoff function $\ell_\pi(s, a)$ of the base game $\mathcal{G}$. The objectives for an average or discounted reward MDP problem (Puterman, 2005) are easily derived from the repeated payoff structures in infinitely repeated games against nature we define on Sec. B.1.

2.3 COMPLETE GAMES AGAINST NATURE

The framework of games against nature that we have described thus far proposes games that are *incomplete* in the sense that nature's player payoff function is constant and unknown. We extend the original formulation to consider a nature player that optimizes for a non-constant payoff function under one of *principles of choice* for decision-making under uncertainty. More formally,

Definition 5 ***(Complete Game Against Nature)*** *A* ***complete game against nature*** *is an extension of a game against nature $\mathcal{G} = \langle \mathcal{S}, \mathcal{A}, \ell_\pi, \ell_\rho \rangle$ where the nature player ρ plays with strategy $\rho : \mathcal{A} \to \Delta(\mathcal{S})$ and maximizes a non-constant payoff $\ell_\rho : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.*

As before, we focus on the principle of *expected utility* principle and define nature's expected payoff to be $\mathcal{L}_\rho = \mathbb{E}_{\substack{a\sim\pi \\ s\sim\rho}}[\ell_\rho(s,a)]$ against the mixed strategy π of the decision-making agent. Repeated nature-agent interactions prompt a repeated game for which equivalent MDP and RL problems emerge for nature's player under a similar expected objective:

$$\mathcal{L}_\rho^t = \mathbb{E}_{p(\tau|\pi,\rho)}\left[\ell_\rho^t(\tau)\right] \tag{2}$$

Holding the assumptions of perfect information and sequential nature-agent interactions as before, a nature player sequential, repeated, perfect information game against the agent induces the following MDP problem:

Definition 6 ***(Nature MDP)*** *A **nature** MDP problem $\mathcal{N} = \langle \mathcal{A}, \mathcal{S}, \pi, r, \pi_0 \rangle$ is an MDP with state space $\mathcal{A}$, action space S, transition dynamics $\pi : \mathcal{S} \to \Delta(A)$, reward function $r : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, and initial state distribution $\pi_0 : \Delta(\mathcal{A})$ is the sequential game a nature agent ρ plays against a decision-making agent π.*

Thus, every complete game induces a pair $(\mathcal{M}, \mathcal{N})$ of dual MDPs (see Fu et al. (2021)). In this case, the initial state distribution π_0 of $\mathcal{N}$ is defined in terms of the initial state distribution ρ_0 of the decision-making agent MDP $\mathcal{M}$ such that $\pi_0(a_0) = \sum_s p_0(s_0)\pi(a_0|s_0)$. Moreover, nature's policy ρ is the transition dynamics $\rho(s_t|s_{t-1}, a_{t-1})$ of the decision-making agent's MDP and, conversely, the decision-making agent policy or strategy π is the transition function $\pi(a_t|s_t)$ of nature's MDP. Also, the reward functions on each MDP are the payoff functions ℓ_ρ and ℓ_π on the complete base game $\mathcal{G}$.

3 Mechanism Induced Games Against Nature

In this section, drawing inspiration from the field of mechanism design (Vickrey, 1961) (see Sec. B.2), we consider the emergence of complete games against nature that assume transformations on the payoff functions of both nature and the decision-making agent. We introduce the figure of *the principal*. This third player can manipulate nature and the decision-making agent payoff structures. Furthermore, motivated by adaptive approaches to mechanism design, we describe *games against the principal*, a three-player game between the decision-making agent, nature, and the principal. In these games, the principal's objective is to discover the optimal outcome function from a restricted set that maximizes either player's total payoff.

3.1 Mechanism Design for Games Against Nature

A mechanism design problem for *games against nature*, a *principal* defines the base game $\mathcal{G}$, both players action spaces $\mathcal{S}$ and $\mathcal{A}$, and their modified outcomes $\tilde{\ell}_\rho$ and $\tilde{\ell}_\pi$. We restrict our interest to cardinal outcomes (e.g., money, years in prison, or scalar reward) and structured outcome spaces $\mathcal{O} = O_1 \times \ldots \times O_n$. We assume that utilities $\ell_i : O_i \to \mathbb{R}$ are of the form $\ell_i = o_i$ where the $i-th$ player maximizes the outcome o_i provided by the mechanism. Under these assumptions, we define the problem of designing a mechanism for nature-agent game as follow:

Definition 7 ***(Mechanism of a Game Against Nature)*** *Let $\mathcal{R} = (r_\rho, r_\pi)$ be the set of possible outcomes for nature and the decision-making agent, the **mechanism of a game against nature** for $\mathcal{R}$ is a tuple $\mathcal{F} = \langle \mathcal{R}, \mathcal{A} \times \mathcal{S}, r \rangle$ where $\mathcal{A}$ and $\mathcal{S}$ are the decision-making and nature player action spaces and $r : \mathcal{A} \times \mathcal{S} \to \mathcal{R}$ is the mechanism outcome function.*

Thus, any mechanism $\mathcal{F}$ defined this way will induce a nature-agent game $\mathcal{G}_\mathcal{F}$ with the following structure:

Definition 8 ***(Mechanism-Induced Game Against Nature)*** *Given a mechanism of a game against nature $\mathcal{F} = \langle \mathcal{R}, \mathcal{A} \times \mathcal{S}, r \rangle$ and the utility functions $\ell_\pi : \mathcal{R} \to \mathbb{R}$ and $\ell_\rho : \mathcal{R} \to \mathbb{R}$, a **mechanism-induced game against nature** is a tuple $\mathcal{G}_\mathcal{F} = \left\langle \mathcal{A} \times \mathcal{S}, \tilde{\ell}_\pi, \tilde{\ell}_\rho \right\rangle$ where $\mathcal{A}$ and $\mathcal{S}$ are the action spaces of the decision-making agent and nature, and the utility functions $\tilde{\ell}_\pi : \mathcal{R} \to \mathbb{R}$ and $\tilde{\ell}_\pi : \mathcal{R} \to \mathbb{R}$ are defined such as:*

$$\tilde{\ell}_\pi(s,a) = \ell_\pi(r(s,a)) = r_\pi \tag{3}$$

$$\tilde{\ell}_\rho(s,a) = \ell_\rho(r(s,a)) = r_\rho \tag{4}$$

for outcomes r_π and r_ρ and the outcome function r of the mechanism $\mathcal{F}$.

This definition assumes that both players have linear utility functions that ignore the opponent's outcome. Furthermore, the notation for the outcome function $r : \mathcal{A} \times \mathcal{S} \to \mathcal{R}$ is intentionally chosen to highlight that, in a mechanism-induced game against nature, the outcome function establishes the payoff (or reward) functions $\tilde{\ell}_\rho$ and $\tilde{\ell}_\pi$ for a stage complete game against nature $\mathcal{G}_\mathcal{F}$ the induced dual MDP problems.

3.1.1 MANUALLY-DESIGNED MECHANISMS

At first, we are interested in mechanisms where the principal is an entity external to the design process. Our interest stems from the connection among the extensive collection of RL problems (Bellemare et al., 2012; Brockman et al., 2016), their underlying MDP, and equivalence to repeated games against nature we introduce in Sec. 2. We reason that, in this context, *problem designers* fulfill the role of the principal. They define the action spaces for both nature and the decision-making agent and the payoff structure of the latter. If one establishes an appropriate payoff function for nature, it is possible to convert every single-agent MDP into a complete game against nature. Inspired by the types of evolutionary pressures nature poses over living agents, in line with existing methods and with the principles of choice of *minimax* and *maximax* (Milnor, 1951; LaValle, 2006), we formulate cooperative and competitive complete agent-nature games.

Definition 9 ***(Adversarial Game Against Nature)*** *An **adversarial game against nature** $\mathcal{G}^- = \langle \mathcal{S}, \mathcal{A}, \ell_\pi, \ell_\rho \rangle$ is a zero-sum game the decision-making agent plays against nature such as nature non-constant maximizes a payoff function $\ell_\rho = -\ell_\pi$, or in other words, the total outcome of the game $\mathcal{G}^-(\pi, \rho) = \ell_\pi + \ell_\rho = 0$. In this adversarial, zero-sum game both player's optimization objectives can be succinctly written down as:*

$$J^-(\pi, \rho) = \min_\rho \max_\pi \ell_\pi \tag{5}$$

On the other hand, we can also define a game where both players cooperate to maximize the payoff function of the decision-making agent:

Definition 10 *A **cooperative game against nature** $\mathcal{G}^+ = \langle \mathcal{S}, \mathcal{A}, \ell_\pi, \ell_\rho \rangle$ is a non-zero game where the decision-making agent plays with nature's player that aims to maximize a non-constant payoff $\ell_\rho = \ell_\pi$. This cooperative game objective can be succinctly written down as:*

$$J^+(\pi, \rho) = \max_\rho \max_\pi \ell_\pi \tag{6}$$

3.1.2 AUTOMATICALLY-DESIGNED MECHANISMS

Motivated by previous adaptive approaches to mechanism design (Vorobeychik et al., 2007; 2012; Baumann et al., 2018), we are interested in incorporating the principal as a player, with strategy ν, a payoff function ℓ_v (see Sec. B.2). We construct a three-player game between the decision-making agent, nature's player, and a principal. More formally,

Definition 11 *A **game against a principal** is a tuple $\mathcal{G}_\nu = \left\langle \mathcal{R}, \mathcal{A}, \mathcal{S}, \ell_\nu, \tilde{\ell}_\pi, \tilde{\ell}_\rho \right\rangle$ that denotes an asymmetric three-player game between a decision-making agent π with strategy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, nature's player $\rho : \mathcal{A} \to \Delta(\mathcal{S})$, and a principal player ν with strategy $r : \mathcal{A} \times \mathcal{S} \to \mathcal{R}$ that can manipulate the other players payoff functions $\tilde{\ell}_\pi$ and $\tilde{\ell}_\rho$ in benefit of maximizing its own payoff function ℓ_ν whose objective is to find the optimal mechanism $r^* \in R$ over a space of feasible mechanisms R, such that:*

$$r^* = \arg\max_{r \in R} \ell_\nu \tag{7}$$

Coalitional Games Similarly, we explore adversarial and cooperative games against a *dictatorial principal* ($\ell_\nu = \min_i \tilde{\ell}_{i=k}$) (Sandholm, 2003) who forms a coalition with either nature or the decision making-agent. From the perspective of the decision-making agent, the core of RL and MDP problems, in a **cooperative nature-principal game**, nature and the principal form a *coalitional*

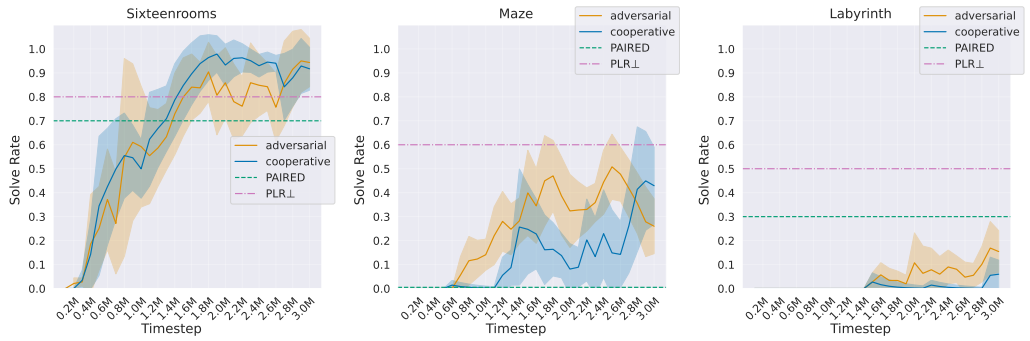

Figure 1: Our exploratory experiments shows that both cooperative and adversarial agent-principal games (Eq. 10 and Eq. 11) offer competitive zero-shot generalization when compared to state-of-the-art methods on *unsupervised environment design* PAIRED (Dennis et al., 2020) and PLR$^{\perp}$ (Jiang et al., 2021) at 250 million steps versus 3 millions in our approach. In *FourRooms*, we significantly outperform both methods with almost two orders of magnitude fewer training samples. The performance on *Maze* and *Labyrinth* degrades, thus requiring further investigation. We describe the full experimental setup in Appendix A

game (Myerson, 1991) where they cooperate to maximize nature's player payoff $\tilde{\ell}_\rho$. We can write down this objective more succinctly as:

$$J(\rho, r) = \max_r \max_\rho \tilde{\ell}_\rho \tag{8}$$

On the other hand, in a **cooperative agent-principal game**, the decision-making agent and the principal player form a coalition that cooperates to maximize the decision-making payoff function $\tilde{\ell}_\pi$ with the joint objective:

$$J(\pi, r) = \max_r \max_\pi \tilde{\ell}_\pi \tag{9}$$

In the context of the three-player game against a principal $\mathcal{G}_\nu$, coalitions formed with the principal provide a strategic advantage for the allied player. For instance, in a cooperative mechanism-induced nature-principal game, if nature's objective is adversarial (Eq. 5), a game against principal would have an objective where the nature-principal coalition minimizes the decision-making agent payoff:

$$J(\rho, \pi, \nu) = \min_{\rho, r} \max_\pi \tilde{\ell}_\pi \tag{10}$$

On the other hand, if the nature-agent game is cooperative, a nature-principal coalition would result in an objective that maximizes the decision-making agent payoff:

$$J(\rho, \pi, \nu) = \max_{\rho, r} \max_\pi \tilde{\ell}_\pi \tag{11}$$

While other combinations of coalitions are possible, we focus in the remainder of this work on the objectives in Eq. 10 and Eq. 11 to show the effects that concentrating the principal power on the nature's player have over generating mechanism-induced game against nature, their equivalent RL and MDP problems, and decision-making adaptation of strategy learned on these coalitional games.

3.2 Preliminary Results

In Appendix A, we include an exploratory ablation experiment that explores the effects that coalitions with a dictatorial principal have on the final zero-shot generalization of an RL agent under cooperative and cooperative objectives. We leverage the experimental setting proposed in the context of unsupervised environment design (Dennis et al., 2020). In this context, we further show (see Figure 1) how the more simplified composition of the objectives in our ablation study leads to comparable, and in some cases, superior zero-shot generalization to existing, more cumbersome state-of-the-art methods (Dennis et al., 2020; Jiang et al., 2021) with almost two orders of magnitude fewer samples.

4 GENERALIZATION GAMES FOR REINFORCEMENT LEARNING

4.1 LEARNING IN GAMES AGAINST NATURE

Players' iterations at playing a game allow them to adapt their strategies to face opponents. In games against nature, a decision-making agent can adapt its strategy against an unknown nature's player whether nature's strategy remains stationary (same player's strategy) throughout their repeated interactions or not. Specifically, in repeated games against nature $\mathcal{G}^t$, the optimal strategy π^* of the decision-making agent should maximize the cumulative payoff functions over t-iterations of the base game, an objective that we can state as:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{p(\tau|\pi,\rho)} \left[\ell^t_\pi(\tau)\right] \tag{12}$$

where the payoff ℓ^t_π is defined as described in Sec. B.1. Note that the objective in Eq. 12 is the traditional RL objective with the exception that we make explicit the dependency on the nature player strategy. Similarly, nature's player optimal strategy in a repeated game against a decision-making agent $\mathcal{G}^t$ is the cumulative payoff function $\ell^t_\rho : (\mathcal{S} \times \mathcal{A})^t \to \mathbb{R}$ similarly selected in consequence to the *finiteness* of interaction length t (see Sec. B.1). Nature's learning objective in complete games (Eq. 2) leads to finding an optimal strategy ρ^* as:

$$\rho^* = \arg\max_{\rho} \mathbb{E}_{p(\tau|\pi,\rho)} \left[\ell^t_\rho(\tau)\right] \tag{13}$$

The objective is equivalent to an RL problem in nature's MDP. A similar argument can be made for the principal agent.

4.2 GENERALIZATION IN GAMES AGAINST NATURE

Assume that the strategies of the nature's player $\rho_\phi : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, the decision-making agent and $\pi_\theta : \mathcal{S} \to \mathcal{A}$ and a principal $r_\omega : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are parametrized by parameters ϕ, θ, ω respectively. Further, consider the consequences rendered over a RL agent by the non-stationarity induced by nature and the principal players learning to optimize their respective payoff functions. Our proposal of a **Generalization Games for Reinforcement Learning** concentrates on the three RL problems structured as repeated non-differentiable games against the principal, where each agent learns and execute a mixed and parametric strategy.

Definition 12 ***(Generalization Game for Reinforcement Learning)*** *A Generalization Game for RL is a repeated non-differentiable optimization game against a principal* $\mathcal{G}^t = \left\langle \mathcal{R}, \mathcal{A}, \mathcal{S}, \ell_\nu, \tilde{\ell}_\pi, \tilde{\ell}_\rho \right\rangle$ *with objective* $J(\phi, \theta, \omega)$ *among the nature player* ρ_ϕ*, the principal* r_ω*, the decision-making agent* π_θ *were each may simultaneously optimize for non-constant cumulative payoff functions such that:*

$$\theta_n = \theta_{n-1} + \eta \nabla_\theta J(\phi, \theta, \omega) \tag{14}$$

$$\phi_k = \phi_{k-1} + \eta \nabla_\phi J(\phi, \theta, \omega) \tag{15}$$

$$\omega_m = \omega_{m-1} + \eta \nabla_\omega J(\phi, \theta, \omega) \tag{16}$$

reflects the update rules equations of each strategy possibly optimized using Monte Carlo gradient estimation techniques (Mohamed et al., 2020).

Notice that the update function of the decision-making agent strategy π_θ is the traditional RL problem. In contrast, the updates rule for nature's strategy ρ_ϕ and that of the principal r_ω introduce the variability over transition dynamics, initial state distribution (nature's initial strategy), and reward functions (the principal strategy). Moreover, an alternative for learning is to induce non-stationarity by drawing samples from distributions $\phi_k \sim p(\phi)$ and $\omega_m \sim p(\omega)$ for nature and the principal strategies (Tobin et al., 2017; Wang et al., 2019). Combined with the adversarial and cooperative objectives we introduced in Sec. B.2, we account for the axis of optimization in Portelas et al. taxonomy. Thus, the non-stationarity of learned or randomized strategies in these games makes the aspects of current generalization methods innately emerge from our formulation.

5 Conclusions and Future Work

We studied the problem of generalization in RL and introduced the *Generalization Games for Reinforcement Learning* by leveraging and expanding the game-theoretic framework of games against nature, turning the traditional MDP formulation into a multiagent problem. Future work will include larger-scale empirical results, including extending our approach to more complex environments possibly with continuous state and actions spaces (Brockman et al., 2016).

References

Dilip Abreu. On the theory of infinitely repeated games with discounting. *Econometrica: journal of the Econometric Society*, 56(2):383–396, 1988.

Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from Multi-Agent autocurricula. In *International Conference on Learning Representations*, 2020.

Tobias Baumann, Thore Graepel, and John Shawe-Taylor. Adaptive mechanism design: Learning to promote cooperation. June 2018.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. July 2012.

Tapan Biswas. *Decision-Making under Uncertainty*. Palgrave, London, 1997.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI gym. June 2016.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. `https://github.com/maximecb/gym-minigrid`, 2018.

Vincent Conitzer and Thomas Sandholm. Complexity of mechanism design. In *UAI*, 2002.

Vincent Conitzer and Tuomas Sandholm. Applications of automated mechanism design. In *UAI*, 2003. Accessed: 2022-2-26.

Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. December 2020.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. pp. 1–14, 2016.

Justin Fu, Andrea Tacchetti, Julien Perolat, and Yoram Bachrach. Evaluating strategic structures in Multi-Agent inverse reinforcement learning. *The journal of artificial intelligence research*, 71: 925–951, August 2021.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780, nov 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL `https://doi.org/10.1162/neco.1997.9.8.1735`.

Minqi Jiang, Michael Dennis, Jack Parker-Holder, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Replay-Guided adversarial environment design. October 2021.

Khimya Khetarpal, M Riemer, I Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *undefined*, 2020.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of generalisation in deep reinforcement learning. November 2021.

Steven M LaValle. *Planning Algorithms*. Cambridge University Press, May 2006.

Joel Z Leibo, Edward Hughes, Marc Lanctot, and Thore Graepel. Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research. *arXiv:1903.00742 [cs, q-bio]*, March 2019.

John Milnor. Games against nature. Technical report, RAND PROJECT AIR FORCE SANTA MONICA CA, 1951.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *Journal of machine learning research: JMLR*, 21(132):1–62, 2020.

Roger B Myerson. Mechanism design by an informed principal. *Econometrica: journal of the Econometric Society*, 51(6):1767–1797, 1983.

Roger B Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, March 1991.

Leslie Pack Kaelbing, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 1996(4):237–285, May 1996.

Christos H Papadimitriou. Games against nature. *Journal of Computer and System Sciences*, 31(2): 288–301, October 1985.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*, 2017.

Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. Automatic curriculum learning for deep RL: A short survey. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, California, July 2020. International Joint Conferences on Artificial Intelligence Organization.

Martin L Puterman. *Markov Decision Processes*. 2005.

Simon Ramstedt and Chris Pal. Real-Time reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Fereshteh Sadeghi and Sergey Levine. CAD2RL: Real Single-Image flight without a single real image. November 2016.

Tuomas Sandholm. Automated mechanism design: A new application area for search algorithms. In *Principles and Practice of Constraint Programming – CP 2003*, Lecture notes in computer science, pp. 19–36. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

L J Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46 (253):55–67, 1951.

Leonard Jimmie Savage. *The Foundations of Statistics*. Wiley, 1954.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL `http://arxiv.org/abs/1707.06347`.

Richard S Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in Neural Information Processing Systems 8*, pp. 1034–1044. MIT Press, 1996.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. March 2017.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press, Princeton, NJ, 1944.

Yevgeniy Vorobeychik, Daniel M Reeves, and Michael P Wellman. Automated mechanism design in infinite games of incomplete information: Framework and applications. *AAAI Spring Symposium: Game*, 2007.

Yevgeniy Vorobeychik, Daniel M Reeves, and Michael P Wellman. Constrained automated mechanism design for infinite games of incomplete information. *Autonomous agents and multi-agent systems*, 25(2):313–351, September 2012.

Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O Stanley. Paired Open-Ended trailblazer (POET): Endlessly generating increasingly complex and diverse learning environments and their solutions. January 2019.

Shimon Whiteson, Brian Tanner, Matthew E Taylor, and Peter Stone. Protecting against evaluation overfitting in empirical reinforcement learning. In *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*. IEEE, April 2011.

A Experiments

A.1 Experimental Goals

Our experiments aim to understand better the principal's influence on environment generation and the subsequent consequences on a learner agent's generalization ability.

Specifically, we are interested in out-of-distribution zero-shot (OODZ) generalization. This metric informs us of the learner agent's capacity to adapt to novel environments, directly linked to the environments presented during training (as confirmed by our experiments). Intuitively, some reward signals lend themselves to vastly more powerful OODZ generalization than others. In the same vein, some reward signals lead to degenerate solutions on nature's side.

We explore a series of reward signals for the nature coalition. We first study the extrinsic rewards traditionally used for the environment generation problem setting. Then, we study rewards derived from an Intrinsic Curiosity Module (ICM) as described in (Pathak et al., 2017) as a way to cut extrinsic rewards out of the equation. Finally, we investigate a mix of ICM signals and extrinsic rewards. Where applicable, we present both adversarial and cooperative coalitions. In every case, the learner agent receives extrinsic rewards.

Finally, as a small investigation into limiting the nature coalition's powers, we present results for which the coalition was not allowed to choose the placement of the maze's walls (or their amount) but was allowed to choose the goal and the agent's placement. Instead, blocks are uniformly sampled.

A.2 Experimental Setup

We evaluate our series of training regimes in the popular environment Minigrid, described in Chevalier-Boisvert et al. (2018). Minigrid is a partially observable maze environment where the decision-making agent must take discrete actions to reach a goal. When the said goal is reached, the agent receives a sparse reward that decreases with the number of steps. The learner's observation is a 7-by-7 pixel image representing the maze directly in front of it, including its current position. Its vision does not penetrate walls. Nature's (and the principal's) observation is the entire grid it is currently building. During training, we utilize 15-by-15 grids, and in them, up to 50 blocks can be placed. Evaluation grid sizes vary by environment.

We train each scheme on three different seeds for 3M steps, except for the Adversarial ICM Mixing scheme, which instead utilizes seven seeds. For each seed, we evaluate the learner agent's zero-shot transfer in a list of challenging OOD environments shown in Figure 5 for ten evaluation runs each. All learning agents are Proximal Policy Optimization agents using an Long Short Term Memory-based recurrent policy (see Schulman et al. (2017) and Hochreiter & Schmidhuber (1997), respectively). We present aggregations of the OODZ generalization performance in the evaluation environment in this section and C.

Additionally, to ensure Minigrid's reward function does not bias our results, we also explore a variation of it. This new reward function gives a negative signal whenever the agent tries to walk into a wall. This densifies the training signal but introduces a new mechanic into training: nature's actions can now create negative signals for the learner agent.

A.3 RESULTS

Results for which blocks were randomly placed are marked with "RB". Results involving the altered Minigrid extrinsic reward signal are marked with "NB".

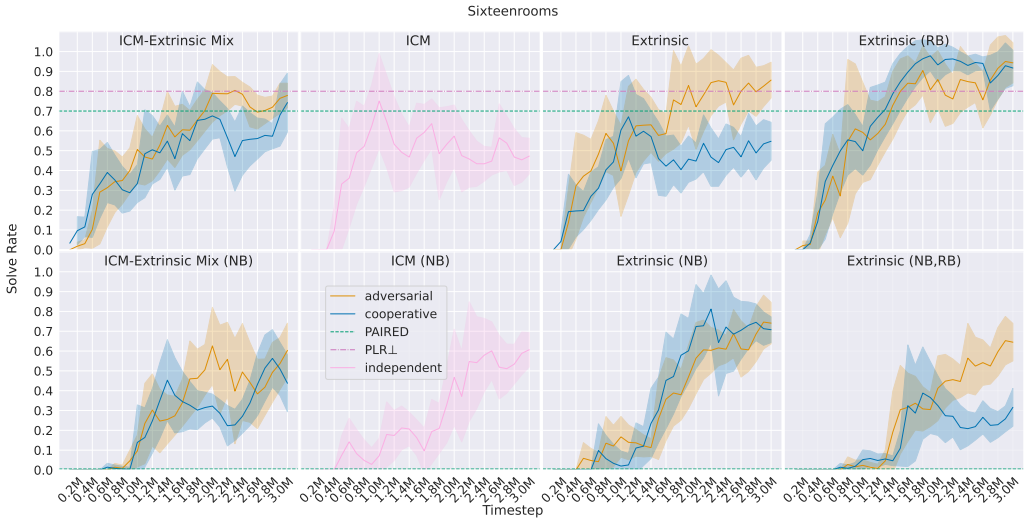

Figure 2: Zero-shot OOD generalization depends greatly the environment generator's reward signal. Here pictured: generalization on the Sixteenroom environment.

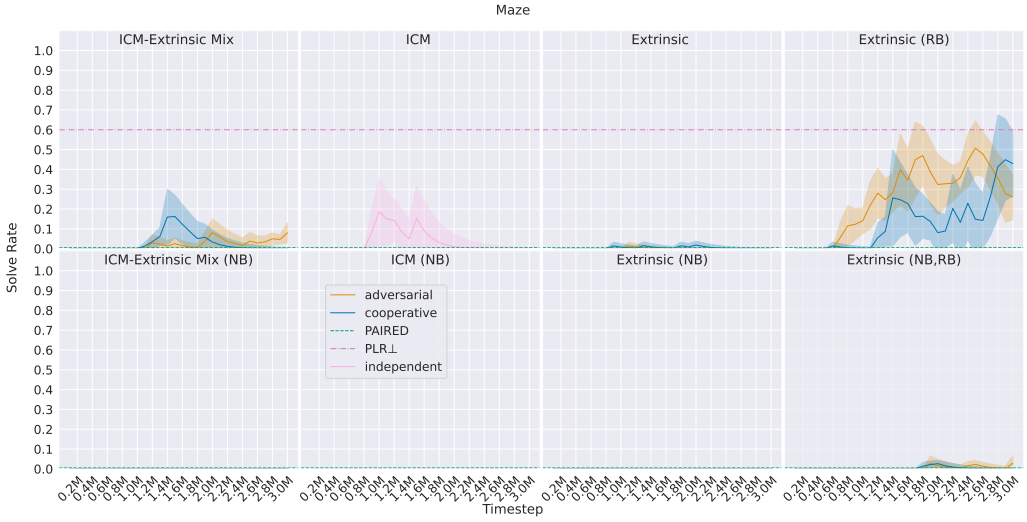

Figure 3: Zero-shot OOD generalization depends greatly the environment generator's reward signal. Here pictured: generalization on the Maze environment.

Our best performing learners overall involve coalitions learning from extrinsic rewards without being able to place blocks. Both in the cooperative and the adversarial setting, they are consistently both the fastest to train and the best at generalization. The exception is when using the altered extrinsic signal: in that case, allowing the coalition to pick the placement of the blocks leads to the best

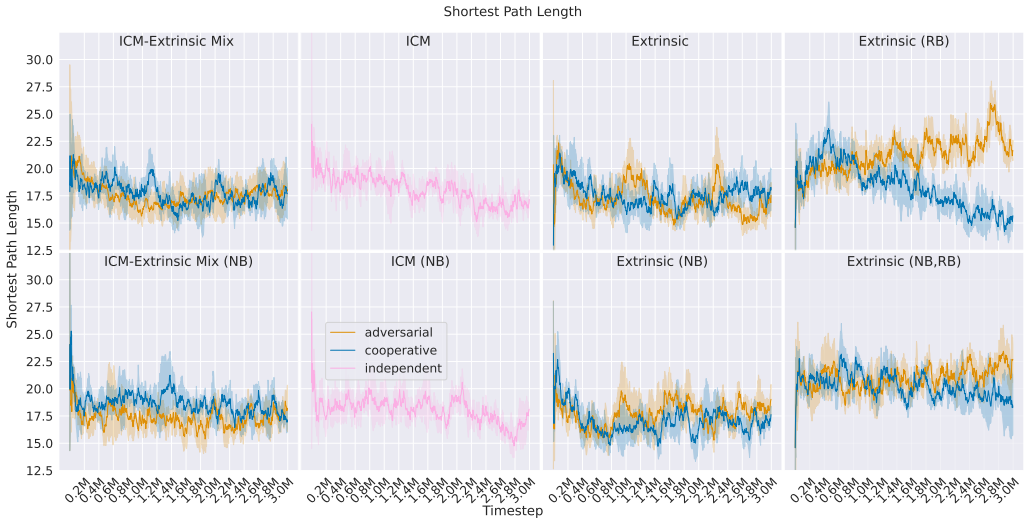

Figure 4: The "Shortest Path Length" is the length of the optimal path from the agent's original position to the goal. Note that this plot uses generous smoothing in order to highlight our findings. It also does not take impassable environments into account.

generalization across evaluation environments. However, the former scheme still performs better in the most complex environments (labyrinth and maze). Preventing the coalition from placing the blocks discouraged it from making trivial (too-hard) environments in the cooperative (adversarial) scheme. This induces more environment diversity for the learner.

As for ICM, interestingly, mixing in extrinsic rewards yields close to the same generalization prowess as using the direct ICM signal, for both the original and the altered extrinsic rewards and the cooperative and adversarial schemes (but in both cases, the adversarial mixing does do better than the cooperative one). This suggests that ICM, or other intrinsic signals, could be used as a replacement for the nature coalition's extrinsic-based training signal. As they are entirely independent of the extrinsic reward signal, these intrinsic objectives could be used to build much more robust training regimes for environment generators.

The relation between adversarial and cooperative schemes is interesting. In almost every scheme and environment, adversarial schemes tend to do slightly better than their counterparts. In the case of a nature coalition with no power over block placement, the reason is apparent: as seen in Figure 4, the paths generated by the adversarial scheme are longer than the cooperatives. This explains the gap in OODZ generalization for that scheme: learners trained in the adversarial scheme are tasked with solving much longer mazes, which directly translates into experiencing more varied environments. For other schemes, the gap is less obvious. Still, the adversarial (cooperative) coalitions find alternative ways to make difficult (easy) environments: consistently, when allowed to choose the amount and placement of blocks, adversarial schemes place more blocks than cooperative schemes. See Figure 10 for more details.

Notice in Figure 2 that many of our training regimes yield decision-making agents that beat the reported OODZ generalization performance of PLR$\perp$ and PAIRED (as described in Jiang et al. (2021) and Dennis et al. (2020), respectively) an order of magnitude earlier than both of those methods on the Sixteenroom environment. Our learners need less than 3M steps to reach 85% solve rate in Sixteenroom, while PLR$\perp$ and PAIRED require 250M to reach about the same. Specifically, the agent trained with a nature coalition that cannot pick blocks (using Minigrid's original extrinsic signal) reaches an almost 95% solve rate in only 1.9M steps. More comparisons of this sort are available in appendix C.

We have also experimented with the altered Minigrid reward function on PAIRED. When PAIRED uses said reward, it produces a learner agent that is utterly incapable of solving SixteenRooms, Maze, and Labyrinth. Its regret formulation explains this: placing a block penalizes the antagonist; thus, the adversary is incentivized to make the simplest tasks possible. The more blocks there are, the more

the antagonist bumps into them and thus, the more negative the reward signal for the adversary. *All* of our schemes do better than PAIRED at this altered reward function, implying that serious thought should be given to extrinsic rewards and their unreliability for environment generation.

B Additional Background

B.1 Payoff Structures in Repeated Games

In repeated games, the decision-making payoff function $\ell_\pi^t(\tau)$ is defined differently depending on *finiteness* of the length or horizon t of repetitions of the base game. In *finite repeated games*, a game $\mathcal{G}^T$ with a finite and bounded number of interactions $T < \infty$, the cumulative payoffs functions $\ell_\pi^T : (\mathcal{S} \times \mathcal{A})^T \to \mathbb{R}$ for can be defined as the sum of individual base game $\mathcal{G}$ payoffs ℓ_π, such that:

$$\ell_\pi^T(\tau) = \sum_{t=0}^{T-1} \ell_\pi(s_t, a_t) \tag{17}$$

However, when the number of interactions is infinite and leads to *infinitely repeated nature-agent games* $\mathcal{G}^\infty$, there are two main choices for payoffs structures. When the decision-making agent is indifferent as to when its one-shot payoffs $\ell_\pi(s_t, a_t)$ is received during the interaction, we can define an *average* payoff (limit of means) such that

$$\ell_\pi^\infty(\tau) = \lim_{T\to\infty} \inf \frac{1}{T} \sum_{t=1}^{T} \ell_\pi(s_t, a_t) \tag{18}$$

Alternatively, if the agent express preferences for one-shot payoffs ℓ_π received earlier on the interaction history, we can introduce a discount factor $0 < \gamma < 1$ (Abreu, 1988) such that:

$$\ell_\pi^\infty(\tau) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \ell_\pi(s_t, a_t) \tag{19}$$

For discount values of $\gamma < 1$, these infinite games are equivalent to a finite game with an effective number of interactions $T = \frac{1}{1-\gamma}$ after which no base game payoff is considered ($\gamma^t = 0$).

B.2 Introduction to Mechanism Design

Mechanism design Vickrey (1961) aim to elicit desirable behavior (e.g., Pareto optimal Nash equilibrium) from the players involved in a game. We introduce the concepts of game form or mechanism and that of the n-player game induced by a mechanism.

Definition 13 *Let $\mathcal{O}$ be a set of possible outcomes, a **game form or mechanism** for $\mathcal{O}$ is a tuple $\mathcal{F} = \langle \mathcal{O}, \mathcal{A}, \varphi \rangle$, where $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_n$ is a cross product of action or strategy spaces for n players and $\varphi : \mathcal{A} \to \mathcal{O}$ is an outcome rule or function.*

Utility functions $\ell_i : \mathcal{O} \to \mathbb{R}$ generally convert player's preference over ordinal outcomes into a cardinal utility values (Von Neumann & Morgenstern, 1944)

Definition 14 *Given a mechanism $\mathcal{F}$ and a set ℓ of utility functions $\ell_i : \mathcal{O} \to \mathbb{R}$ for each player, the n-player **game** $\mathcal{G}_\mathcal{F}$ **induced by the mechanism** $\mathcal{F}$ is a pair $\mathcal{G}_\mathcal{F} = \langle \mathcal{A}, \ell_\varphi \rangle$ where $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_n$ is a cross product of action or strategy spaces for n players and ℓ_φ is the set of modefied utility functions $\tilde{\ell}_i$ for each player, such as:*

$$\tilde{\ell}_i(a_1, \ldots, a_n) = \ell_i(\varphi(a_1, \ldots, a_n)) \tag{20}$$

defines the payoff received by the i-th *agent under a joint strategy $(a_1, \ldots, a_n) \in \mathcal{A}$ defined by the outcome function $\varphi(a_1, \ldots, a_n)$.*

Automated Mechanism Design. The field of automated mechanism design approaches design of the mechanism outcome function $\varphi : \mathcal{A} \to \mathcal{O}$ manually or automated. The approach to automated mechanism design (Conitzer & Sandholm, 2002; Sandholm, 2003) leverages the figure of the principal or designer, a player ν whose objective $\ell_\nu : \Phi \times \ell_\varphi \to \mathbb{R}$ is defined over the space of outcome

functions $\varphi \in \Phi$ and players modified utilities ℓ_φ. Possible objective functions for the principal ν to maximize may include *social welfare (utilitarian)* $\ell_\nu = \sum_i \tilde{\ell}_i$, the *worst utility (egalitarian)* of any player $\ell_\nu = \min_i \tilde{\ell}_i$ or *dictatorial* $\ell_\nu = \tilde{\ell}_{i=k}$, among others (Sandholm, 2003). The principal's objective is selected according to an equilibrium concept required for players in the induced game (Conitzer & Sandholm, 2003). In general, the automated design of a mechanism is reduced to finding, after the outcome and action space are manually defined, the optimal outcome function φ^*, such that the objective:

$$\varphi^* = \arg\max_{\varphi \in \Phi} \ell_\nu \tag{21}$$

enforces the preferences of the principal player ν over the game $\mathcal{G}_F$ induced by the mechanism $\mathcal{F}$.

C Additional Plots

Here are presented the rest of the zero-shot OODZ generalization plots. Results for which blocks were randomly placed are marked with "RB". Results involving the altered Minigrid extrinsic reward signal are marked with "NB".

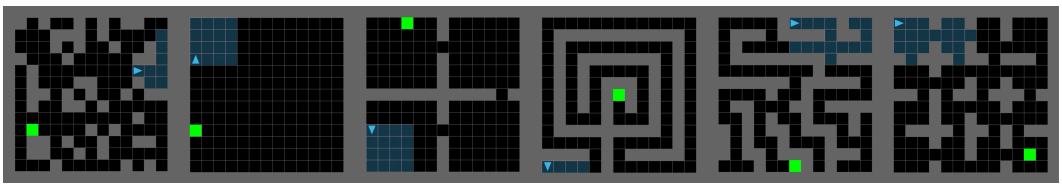

Figure 5: Evaluation environments. From left to right: Cluttered50, Empty-Random-15X15, Four-Rooms, Labyrinth, Maze, SixteenRooms.

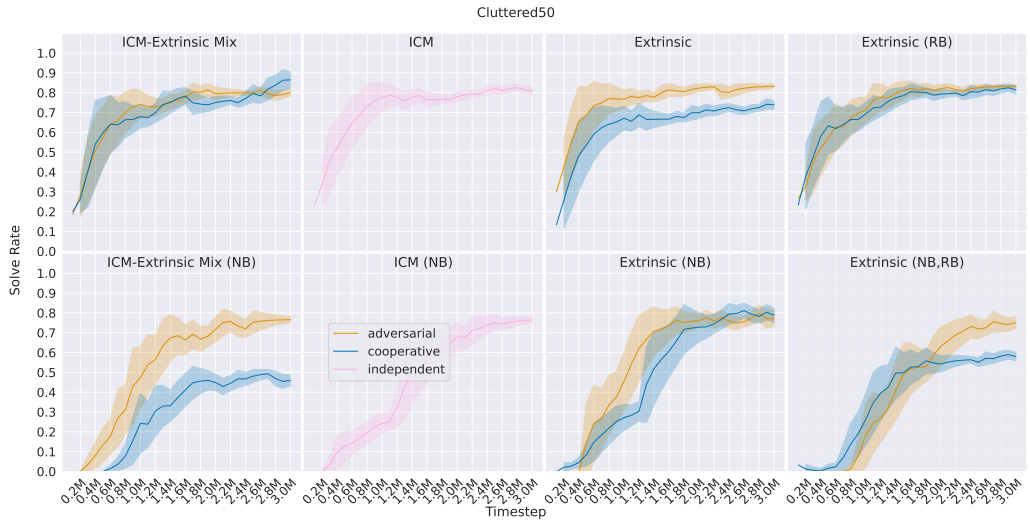

Figure 6: Zero-shot OOD generalization in the Cluttered50 environment.

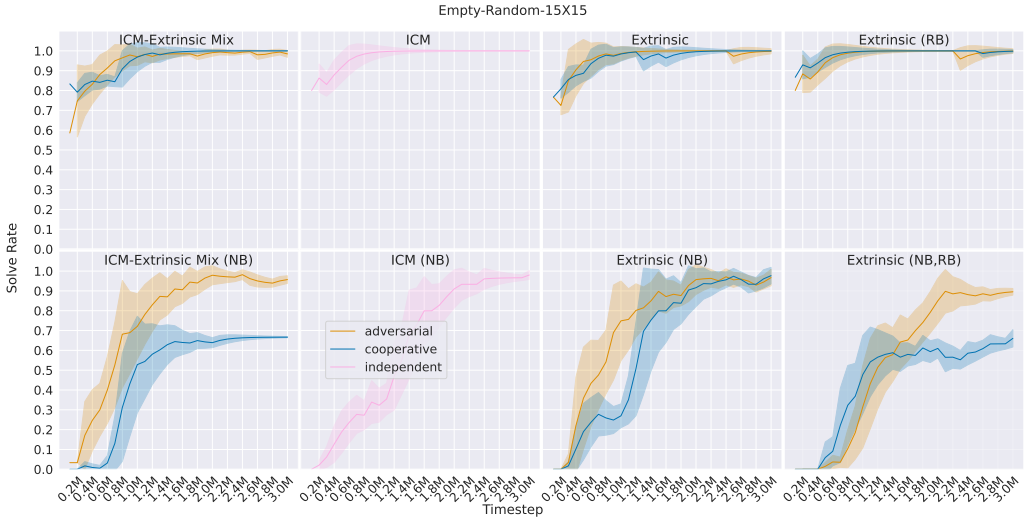

Figure 7: Zero-shot OOD generalization in the Empty-Random-15x15 environment.

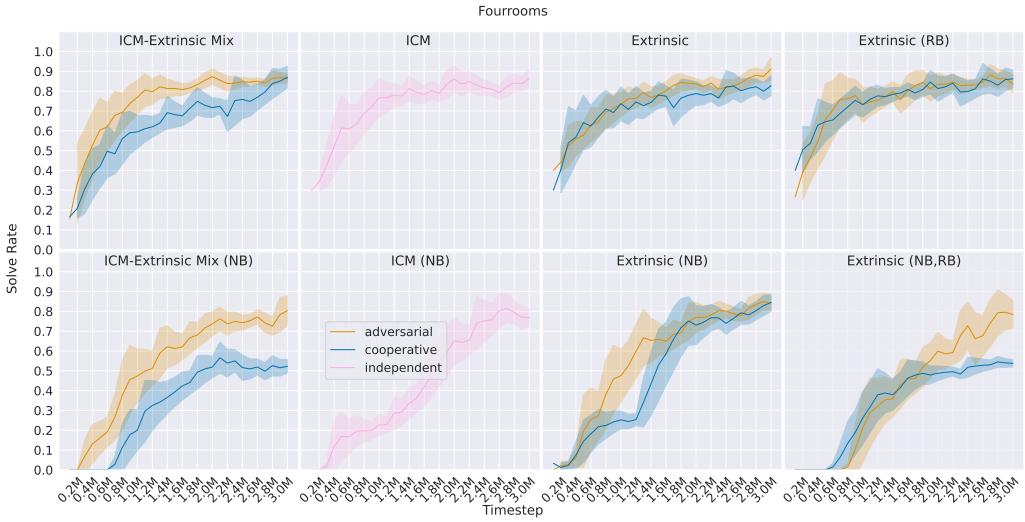

Figure 8: Zero-shot OOD generalization in the FourRooms environment.

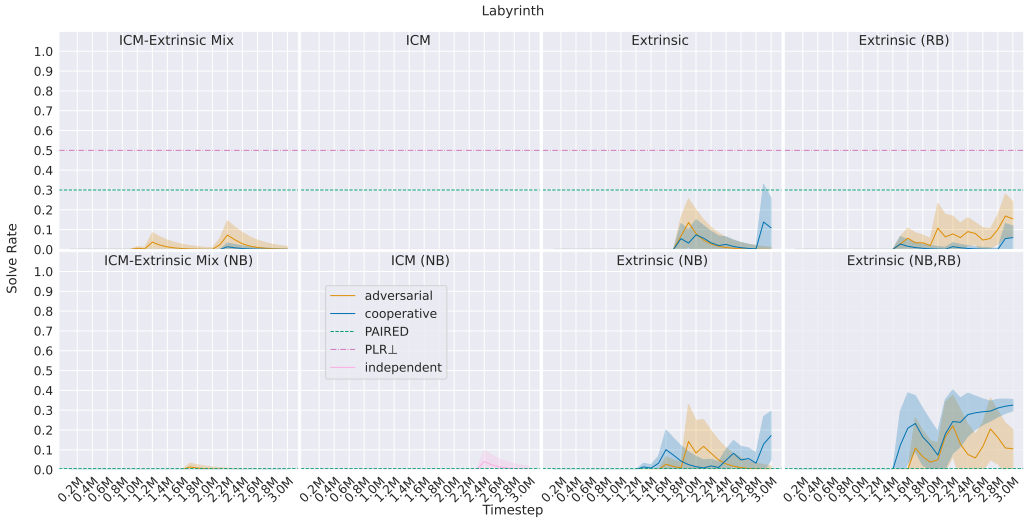

Figure 9: Zero-shot OOD generalization in the Labyrinth environment.

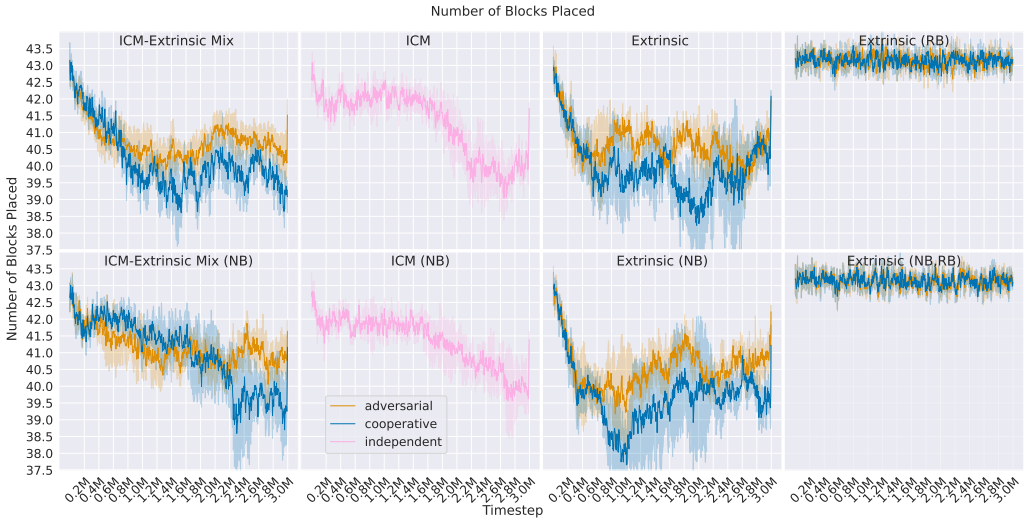

Figure 10: The amount of blocks placed during environment generation environment generation during training. Each data point is one environment. RB schemes place an almost-constant amount since the amount and location of the blocks is fully stochastic.