Multi-agent Reinforcement Learning for Bargaining under Risk and Asymmetric Information

Kyrill Schmid[1], Lenz Belzner[2], Thomy Phan[1], Thomas Gabor[1] and Claudia Linnhoff-Popien[1]
[1]*Distributed and Mobile Systems Group, LMU Munich, Germany*
[2]*MaibornWolff, Germany*
{*kyrill.schmid, thomy.phan, thomas.gabor, linnhoff*}*@ifi.lmu.de, lenz.belzner@maibornwolff.de*

Keywords: Multi-agent Reinforcement Learning.

Abstract: In cooperative game theory bargaining games refer to situations where players can agree to any one of a variety of outcomes but there is a conflict on which specific outcome to choose. However, the players cannot impose a specific outcome on others and if no agreement is reached all players receive a predetermined status quo outcome. Bargaining games have been studied from a variety of fields, including game theory, economics, psychology and simulation based methods like genetic algorithms. In this work we extend the analysis by means of deep multi-agent reinforcement learning (MARL). To study the dynamics of bargaining with reinforcement learning we propose two different bargaining environments which display the following situations: in the first domain two agents have to agree on the division of an asset, e.g., the division of a fixed amount of money between each other. The second domain models a seller-buyer scenario in which agents must agree on a price for a product. We empirically demonstrate that the bargaining result under MARL is influenced by agents' risk-aversion as well as information asymmetry between agents.

1 INTRODUCTION

Bargaining has been stated one of the most fundamental economic activities and describes situations where a group of individuals faces a set of possible outcomes with the chance to agree on an outcome for everyone's benefit (Nash Jr, 1950; Roth, 2012). If the group fails to reach an agreement the participants get their predefined status quo outcomes. A bargaining game for two players is displayed in Figure 1 where the gray area represents the set of feasible outcomes. In this case player 1 wants the actual agreement point to be as far to the right as possible whereas player 2 desires it as high as possible. The actual outcome is the subject of the negotiation between the participants. Importantly, each participant has the ability to veto any agreement besides the status quo outcome.

In the game-theoretic context the analysis of bargaining is commonly based on the assumption of perfectly rational agents. In this line, in (Nash Jr, 1950) an axiomatic model is considered that allows to derive a unique solution for the problem of dividing a common good between two bargainers. Approaches from the fields of evolutionary game theory (Fatima et al., 2005), genetic algorithms (Fatima et al., 2003) and neural networks (Papaioannou et al., 2008) drop the assumption of perfect rationality by introducing boundedly rational agents and analyze the evolution of stable strategies.

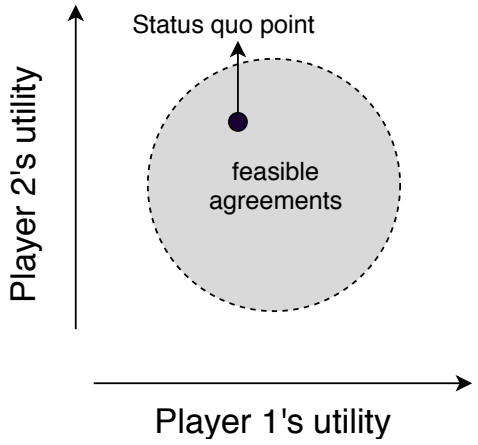


Figure 1: A bargaining game between two players.

In this work the approach of boundedly rational agents is extended by studying bargaining as a multi-agent reinforcement learning (MARL) problem. Specifically, we use MARL to analyze the bargaining outcomes found by two independent bargainers. For this purpose we propose two bargaining environments resembling scenarios that are commonly

encountered in the bargaining literature. The first domain, called divide-the-grid, considers the situation of two bargainers trying to divide a common good that is available in a fixed quantity, e.g., the division of a fixed amount of money. Secondly, we consider a seller-buyer environment which models a negotiation process between a single seller and a single buyer which is also referred as a bilateral monopoly. Though both domains confront the learners with the challenge of finding an agreement, there are differences in the environments with respect to the symmetry of the task, the available information, the payoffs and the representation of the bargaining process. In this way, the environments while sharing the same interface can be used to highlight different aspects of MARL bargaining. On the basis of the proposed domains we conduct two experiments to analyze critical aspects of the MARL bargaining process. Specifically, in this work we make the following contributions:

- We formulate bilateral bargaining as a multi-agent reinforcement learning task by proposing two bargaining domains with a common MARL interface applicable with state of the art RL methods.
- To analyze the influence of risk-aversion on the bargaining outcome we build risk-averse agents on the basis of the *Categorical DQN* (Bellemare et al., 2017) architecture. For risk-averse decision making we use different percentiles from the value distribution instead of the expected value which is normally used for value based methods. We further empirically demonstrate that an agent's risk-aversion negatively affects its bargaining share.
- The second aspect of the analysis is the influence of information asymmetry. Information asymmetry comes in the shape of uncertainty about the quality of a product for which the agents are bargaining a price. In this experiment agent 1 (seller) receives more information on the product quality than agent 2 (buyer). Here it is shown that the bargaining success negatively correlates with the amount of uncertainty.

Both experimental results, though consistent with theoretic results, have to the best of our knowledge not been shown earlier in the case of MARL.

2 BACKGROUND

Nash bargaining describes situations where individuals can collaborate but have to decide in which way to actual collaborate with each other (Nash Jr, 1950). More formally, a bargaining situation is characterized by a set of bargainers N, an agreement set A, a disagreement set D and for each bargainer $i \in N$ a utility function $u_i : A \cup D \to \mathbb{R}$ that is unique up to a positive affine transformation. The set of possible agreements, the disagreement set and the utility functions are sufficient to construct the set S of all utility pairs which are possible outcomes from the bargaining, i.e., $S = \{(u_1(a), .., u_n(a))\}$, and the unique point of disagreement $d = \{u_1(d), ..., u_n(d)\}$. The pair (S,d) is a bargaining problem and builds the model for the axiomatic solution (Nash Jr, 1950).The set of all bargaining problems is denoted B. A bargaining solution is a function $f : B \to \mathbb{R}^2$ that assigns to each bargaining problem $(S,d) \in B$ a unique element of S. Nash stated four axioms which should hold for a bargaining solution (Nash Jr, 1950):

1. *Pareto-Efficiency:* The players will never agree on an outcome if there is another outcome available in which both players are better off.
2. *Invariance to Equivalent Utility Representations:* The bargaining outcome is invariant if the utility function and the status quo point are scaled by a linear transformation.
3. *Symmetry:* It is assumed that all bargainers have the same bargaining ability, i.e., asymmetries between the players should be modeled in (S,d). Therefore, a symmetric bargaining situation should result in equal outcomes for all players.
4. *Independence of Irrelevant Alternatives:* If a bargaining solution found for a bargaining situation can be assigned to a smaller subset then the solution will be the same if the new feasible set is reduced to this subset.

It has been shown that under these axioms there is a unique bargaining solution $f : B \to \mathbb{R}$, given by:

$$f(S,d) = \underset{(d_1,d_2) \leq (s_1,s_2 \in S)}{\arg\max} (s_1 - d_1) \times (s_2 - d_2)$$

Multi-agent Reinforcement Learning. Reinforcement learning (RL) describes methods where an agent learns strategies, also called policies, through trial-and-error interaction with an unknown environment (Sutton and Barto, 2018). Although RL assumes a single agent it has been used in the multi-agent case (Littman, 1994; Tan, 1993) which is known as multi-agent reinforcement learning (MARL). In MARL different challenges arise: the exponential growth of the discrete state-action space, the nonstationarity of the learning problem and an exacerbated exploration-exploitation trade-off (Buşoniu et al., 2010). In spite

of these challenges MARL has been successfully used in a variety of applications and more recently also been extended to methods featuring deep neural networks as function approximators (Nguyen et al., 2018).

In the presence of multiple decision makers the learning problem can be formally described as a Markov game $\mathcal{M}$ which is a tuple $(\mathcal{D}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ consisting of a set of agents $\mathcal{D} = \{1, ..., \mathcal{N}\}$, a set of states $\mathcal{S}$, the set of joint actions $\mathcal{A} = \mathcal{A}_1 \times ... \times \mathcal{A}_{\mathcal{N}}$, a transition function $\mathcal{P}(s_{t+1}|s_t, a_t)$ and a reward function $\mathcal{R}(s_t, a_t) \in \mathbb{R}^{\mathcal{N}}$ (Boutilier, 1996).

So called independent learning refers to methods where agents have no knowledge of other learners. In this case, the goal of an agent is to find a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, that maximizes the expected, discounted return: $G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$ for a discount factor $0 \geq \gamma \geq 1$. One way to find an optimal policy is to learn the action value function $Q_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and use this function as a policy by selecting actions according to a strategy that balances exploration and exploitation during the learning process, e.g., ε-greedy action selection:

$$\pi(s) = \begin{cases} \arg\max_{a \in \mathcal{A}} Q(s, a) & \text{with probability } 1 - \varepsilon \\ \mathcal{U}(\mathcal{A}) & \text{with probability } \varepsilon \end{cases}$$

where $\mathcal{U}(\mathcal{A})$ denotes a random sample from $\mathcal{A}$. At each step an agent i updates its policy with a stored batch of transitions $\{(s, a, r_i, s')_t : t = 1, ...T\}$ and for a given learning rate α by applying the following update rule:

$$Q_i(s, a) \leftarrow Q_i(s, a) + \alpha[r_i + \gamma \max_{a' \in \mathcal{A}_i} Q_i(s', a') - Q_i(s, a)]$$

For deep multi-agent reinforcement learning, each agent i can be represented as a deep Q-Network (DQN) that approximates the optimal action value function.

Risk-averse Learning. The criterion which is commonly used in RL methods for decision making is the expected return, i.e., a policy is optimal if it has the maximum expected return. Risk-sensitive RL describes methods where other criteria are used that also have a notion of risk (Garcıa and Fernández, 2015).

For the purpose of building risk-averse agents we use the *Categorical DQN* model as suggested in (Bellemare et al., 2017) which is build on the DQN architecture but outputs the whole value distribution $p_i(s, a)$ as a discrete probability distribution. In contrast to ε-greedy action selection over the expected action-values we use a policy that selects actions according to their empirical Conditional Value at Risk (*CVaR*) a prominent risk-metric known from portfolio optimization. The definition of *CVaR* is build upon another risk metric which is Value at Risk (*VaR*). Formally, VaR_α for a random variable X representing loss and a confidence parameter $0 < \alpha < 1$, is defined as follows (Kisiala, 2015): $VaR_\alpha(X) := \min\{c : P(X \leq c) \geq \alpha\}$ and can be interpreted as the minimum loss in the $1 - \alpha \times 100\%$ worst cases.

One known drawback of *VaR* is that it does not provide any information for the outcomes in the $1 - \alpha \times 100\%$ worst cases, which might be essential if the potential losses are large. An extension that provides such information is Conditional Value at Risk *CVaR*. For a continuous random variable representing loss and a given parameter $0 < \alpha < 1$, the $CVaR_\alpha$ of X is (Kisiala, 2015):

$$CVaR_\alpha(X) := \mathbb{E}[X | X \geq VaR_\alpha(X)]$$

We build risk-averse agents by making action selection on the basis of the corresponding $CVaR_\alpha$ estimated from the value distribution. Risk-sensitive decision making then means to favor actions with better *CVaR* for a given α. By increasing the confidence parameter α an ever smaller share of worst cases is considered making action selection more sensitive to potential bad outcomes. For the process of exploration a normal $\varepsilon-$greedy selection rule is used.

3 BARGAINING ENVIRONMENTS

In this section, two bargaining environment are introduced for bilateral bargaining. The bargaining scenarios are represented as grids where each configuration of the grid specifies a specific bargaining outcome. To control the bargaining process agents move through the grid by applying their actions $a \in \mathcal{A}$. Through this formulations, the bargaining process is specified through the trajectory of actions during an episode. The domains differ with respect to the task symmetry, i.e., the first task is symmetric meaning both agents receive the same observations and possible payoffs. The second domain in contrast is asymmetric with respect to the available observations and payoffs as agents are either the seller or the buyer of a product.

Divide-the-grid. In this domain we consider the problem of dividing a fixed amount of a discretely divisible commodity, e.g., the the division of a dollar between two agents where the resolution of the division is restricted by the available tokens. Naturally, each agent prefers a higher share to a lower. The commodity is represented through the area of the grid and the two players (red and blue) can claim a share of the commodity by stepping over the cells (Figure

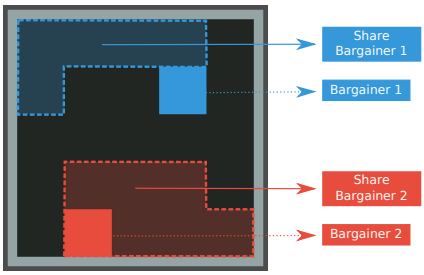


Figure 2: Divide-the-grid models the situation of two bargainers trying to agree on the division of a common good with a fixed quantity, e.g., a dollar. To claim their share, bargainers can step over grid-cells thereby marking it with their respective colors. At the end of an episode, the bargaining shares are given as proportion of the occupied cells for each agent. The bargaining process is abandoned whenever an agent claims a cell that has already been taken by its opponent which gives the disagreement outcomes (d_1, d_2).

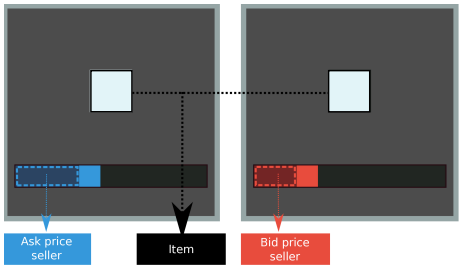


Figure 3: The Seller-buyer environment is a bargaining game with asymmetric information. Agent 1 (blue, left) is the seller of a rectangular item which occurs in two categories, i.e., low quality (dark) or high quality (bright). However, in the case of information asymmetry this information is exclusively available to the seller. The buyer in contrast then receives no information about the object's quality as it merely observes the items shape without information on the object's color.

2). At the end of an episode the bargaining outcome is determined by the number of cells each agent has stepped over. Therefore, if N is the number of cells in the grid and player i has gathered n_i cells during an episode, then the player's share is $\frac{n_i}{N}$. At each episode the bargaining process starts anew, so agents cannot accumulate wealth as for example in multi-stage bargaining processes. The bargaining fails, i.e, results with the status quo point if an agent steps upon cells which have already been claimed by the other agent. In this case, the episode ends and each agent i receives its disagreement reward $d_i = -1 \forall i$. Otherwise, an episode ends after T steps.

Seller-buyer. The second environment resembles a seller-buyer scenario with a single seller and single buyer. This situation is known in economics as a bilateral monopoly. Here, the bargaining problem is to find a price for different products. The seller's task is to set an ask price p_s which is the lowest price the seller is willing to accept for its product. The buyer on the contrary has to give a bid price p_b which is the highest acceptable price to pay. The players can set their prices by sidestepping horizontal through the price field (see Figure 3). The status quo point is realized if the ask price is higher than the bid price in which case the selling was unsuccessful, i.e., $p_b < p_s$. After T steps the bargaining stops and in case of agreement the price is given as the bid price. To model information asymmetry between the players the product at question occurs in different qualities. With perfect information both players would get all information about the current product's quality. However, under asymmetric information only the seller is fully informed. The buyer in this case receives no information on the product quality.

4 RELATED WORK

Most of the work dealing with the study of bargaining comes from the field of game theory where it is commonly assumed that players are perfectly rational. The equilibrium in these models can be found through theoretical analysis of the game and the participants will always play the dominant strategy. In evolutionary game theory the the perfect rationality assumption has been discarded and replaced by allowing agents to be boundedly rational. There is a strong link between evolutionary game theory and multi-agent reinforcement learning (Tuyls and Nowé, 2005) which is why it is considered related work.

Evolutionary Models. In evolutionary game theory (EGT) the assumption of perfect rationality is dropped. Instead, it relies on the concept of populations of different strategies being matched with each other. This leads to an evolutionary process in which strategies with higher relative fitness prevail. Within the framework of EGT bargaining has been studied in a broad range of work. In (Ellingsen, 1997) the evolutionary stability of two classes of strategies is examined. Agents either are obstinate, i.e., their demands are independent of the opponent, or sophisticated agents who adapt to their opponent's expected play. The results show that evolutionary stability of obstinate and sophisticated strategies depends on the certainty of the pie size, i.e., when the pie size is certain evolution favors obstinate agents.

More recently (Konrad and Morath, 2016) considered the effect of incomplete information in a seller (informed) and buyer (uninformed) scenario where

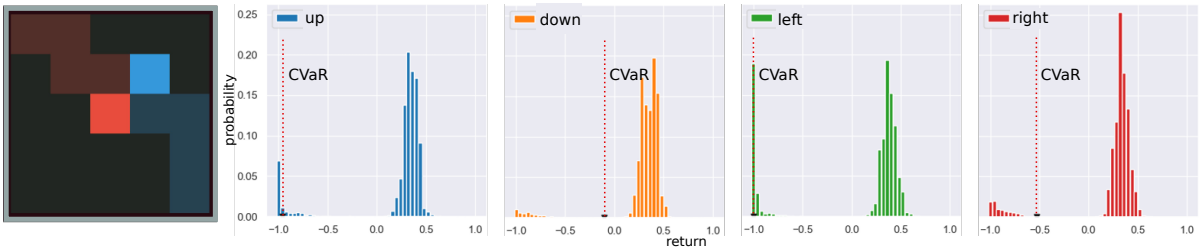


Figure 4: Shown is a snapshot of divide-the-grid (left) and the four resulting histograms representing the value distributions for actions *up, down, left, right* shown in this case for the blue agent. The value distribution associated with the action *left* (third histogram) shows a high probability for a return similar to the disagreement outcome (-1) which is also reflected in a significantly lower *CVaR* value compared to the three other actions.

it is shown that trade becomes less likely when the players apply evolutionary stable strategies compared with the corresponding perfect Bayesian equilibrium. The evolution of fairness is considered in (Rand et al., 2013). The authors demonstrate that when agents make mistakes when judging the payoffs and strategies of others then natural selection favors fairness in the one-shot anonymous ultimatum game.

In (Fatima et al., 2003) study the competitive coevolution in a setting with incomplete information. The authors use a seller-buyer scenario and compare their results with those prescribed from game-theoretic analysis. It is shown, that stable state found by genetic algorithms does not always match the game-theoretic equilibrium. Moreover, the stable outcome depends on the initial population as the players mutually adapt to each other's strategy.

In contrast to approaches from evolutionary game theory in this work we do not consider the evolutionary dynamics of strategies. Rather we study the emergent outcome patterns that result from agent specific parameters such as risk-aversion or environmental modifications such as quality uncertainty. However, we do not model theses parameters as an element of the learning process or test the stability of different populations of strategies against each other.

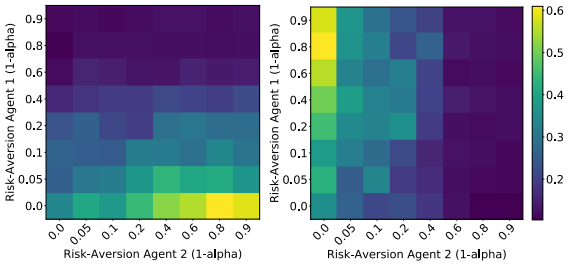
Opponent Modeling. Bargaining between agents is typically a game with incomplete information as an agent normally has no prior knowledge about its opponent strategy. If provided with information about an opponent's preferences or wishes it would be easier for a bidding agent to make an adjusted bid thereby allowing potentially earlier agreements. The building of models from other agents is referred to as *opponent modeling*. With regard to negotiation settings the three main aspects of modeling are: preference estimation (what does the opponent want?), strategy prediction (what will the opponent do?), and opponent classification (what type of player is the opponent?) (Baarslag et al., 2016).

Approaches to build opponent models for negotiation in multi-agent settings come from different fields including Bayesian learning (Gwak and Sim, 2010), non-linear regression (Haberland et al., 2012), genetic algorithms (Papaioannou et al., 2008) and artificial neural networks (Fang et al., 2008). In contrast to approaches from the field of opponent modeling, in this work we do not explicitly build a model of the opponent. Rather an agent considers other agents as a part of the environment which renders the learning problem non-stationary from an individual agents' perspective. In our case adaptation to opponent behavior results from the changing data distribution an agent holds in its sequential memory which used to train its policy.

Social Dilemmas. Recently, deep multi-agent reinforcement learning has been used to study the outcomes of distributed learning in sequential social dilemma domains (Leibo et al., 2017; Lerer and Peysakhovich, 2017; Wang et al., 2018). Social dilemmas are challenging for independent learners as each agent has incentives to make collectively undesirable decisions. In contrast to social dilemmas in bargaining an agent has no incentive to defect for short term personal gain. Rather each agent benefits from an agreement. However, there is a multitude of possible agreements which renders the difficulty of choosing one specific solution.

5 EXPERIMENTS

The experiments described in this section examine the negotiation results learned by independent multi-agent deep reinforcement learning. The first experiment aims at demonstrating how the bargaining outcome depends on the bargainers risk-aversion. The second experiment analyzes how information asymmetry, another important feature for a bargaining process, presses the likelihood of successful bargaining.



(a) Bargaining share agent 1 (b) Bargaining share agent 2

Figure 5: The bargaining result is influenced by agents' risk-aversion. Shown are the bargaining shares for agent 1 (left) and agent 2 (right) for varying levels of risk-aversion modeled through a changing level of α for $CVaR_\alpha$. The share an agent receives in the divide-the-grid setting is dependent on both agent's risk attitude. I.e., when an agent becomes risk-averse, its share is likely to decrease.

The simulations show that risk-aversion and information asymmetry are vital aspects for MARL bargaining.

Risk-aversion. Uncertainty arises in divide-the-grid as the players can not reliably predict the antagonist's behavior. Both players are enabled to stop the bargaining process yielding the less desirable disagreement rewards. In the axiomatic bargaining theory the player's preference orderings are assumed to satisfy the assumptions of von Neumann and Morgenstern. In the case of two risk-averse players both utility functions u_i are concave. A central finding from axiomatic bargaining theory is that whenever one player becomes more risk-averse, then the other player's bargaining share increases (Roth, 2012).

For the purpose of modeling risk-aversion we make use of agents to learn the distribution of the random return as proposed in (Bellemare et al., 2017). In contrast to the common approach of learning policies with respect to the expectation of the return, the availability of the whole value distribution allows to use other metrics for decision making that can be derived from this distribution. More specifically, the value distribution is modeled by a discrete distribution parametrized by $N \in \mathbb{N}$ and $V_{\text{MIN}}, V_{\text{MAX}} \in \mathbb{R}$. The support for this distribution is given by the set of atoms $\{z_i = V_{\text{MIN}} + i\triangle z : 0 \leq i < N\}, \triangle z := \frac{V_{\text{MAX}} - V_{\text{MIN}}}{N-1}$ and the atom probabilities are given by a parametric model $\theta : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^N$ such that the probability for z_i for state s and action a is (Bellemare et al., 2017):

$$p_i(s,a) := \frac{e^{\theta_i(s,a)}}{\sum_j e^{\theta_j(s,a)}}$$

The typically learned value distribution for all actions in the divide-the-grid domain are shown in Figure 2 in this case the value distributions for the blue agent are given. The value distributions reflect the circumstance that choosing *left* in this situation will end the episode and provide the disagreement reward -1 for both agents. The distributions from the other actions in contrast, are less likely to end the episode immediately and therefore also have a higher *CVaR* value, i.e., they have lower associated risk.

In the first experiment we study the effect of an agent's risk-aversion on the bargaining share it receives. The agent's risk-aversion stems from a differing share of worst cases $(1-\alpha)$ which are taken into account. Figures 5a and 5b show the bargaining share for agent 1 and 2 as functions of agent 1's risk-aversion (y-axis) and agent 2's risk-aversion (x-axis) after training for 120.000 episodes. Results are the averages from the last 500 episodes and. The bargaining share for an agent increases with the the other agent's risk-aversion and with its own risk-neutrality.

Asymmetric Information. Bargaining under asymmetric information describes situations where single individuals have more information than others. The aspect of unevenly spread information has been stated as one of the most fundamental difficulties to efficiently coordinate economic activity (Hayek, 1945) and has been extensively studied in economics (Akerlof, 1978; Samuelson, 1984; Kennan and Wilson, 1993). In (Akerlof, 1978) the author demonstrated with the market for "lemons" how the presence of bad quality products for sale has the potential to drive out good quality products and potentially can lead to a market collapse. One form of information asymmetry arises through the existence of goods in different quality grades known as quality uncertainty. For many bargaining situations it is realistic to assume that the seller of the good at question has more or better information on a trading item than the potential buyer.

In this experiment the impact of quality uncertainty on the bargaining success of a bilateral monopoly is examined. The seller and the buyer are independent decision makers both represented as *DQNs*. Quality uncertainty is introduced by allowing the trading object to occur in two different categories, i.e., in low quality and in high quality where the probabilities for the categories are p_l and p_h respectively ($p_h = 1 - p_l$). The seller's payoffs, denoted $\mathcal{U}_s^l, \mathcal{U}_s^h$, are assumed to be strictly larger than the buyer's payoffs, $\mathcal{U}_b^l, \mathcal{U}_b^h$, so that successful bargaining is always preferable as both agents can benefit. The bargaining is successful when at the end of an episode the buyer's bid price is at least as high as the seller's ask price yielding payoffs as given in Table 1. If unsuccessful, the seller receives the payoff equally to its item valu-

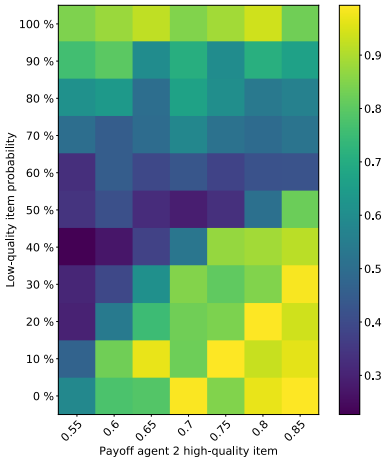


Figure 6: The success of bargaining in a bilateral monopoly with low-quality and and high-quality items is influenced by the probability for low-quality and high-quality to occur and by the buyer's relative valuation of the item.

ation, the buyer in contrast, receives a payoff of 0.

Table 1: Payoffs for the seller and buyer for items with low-quality and high-quality.

	Payoff seller	Payoff buyer
low-quality	0.1	0.3
high-quality	0.4	0.6

To estimate the effect of quality uncertainty on the bargaining success we varied the product quality probabilities p_l, p_h. In the limit the product will only be available in one quality category as $p_l \to 0$ or $p_l \to 1$. In a setting where the product can occur with different qualities the available information for both parties will be critical for the success of the bargaining. Although, for any product category a successful bargaining agreement would make both parties better off the quality uncertainty increasingly rules out this possibility. I.e., if the buying party has less information on the current product its maximum bid price will be derived from its valuation of the low quality product in any case. The better informed seller however, is not willing to sell a high quality product under these terms and will only offer the low quality product on his part. The effect of this information gap is likely to also depend on the relative product valuations of both parties. To control this effect different values for the buyer's high-quality valuations were tested.

Figure 6 shows the results gathered from 12 independent runs where the bargaining success represents the average from the last 100 episodes with a total training time of 120.000 episodes and each episode lasting 11 steps. The results suggest that a successful bargaining becomes less likely for an even ratio of item probabilities. This effect is more severe when the buyer has a very low valuation, i.e., the difference between its payoffs are little. When the buyer's high-quality valuation increases the success rate also increases but is positively skewed towards settings with a higher proportion of high-quality items.

6 CONCLUSION

In this work we studied MARL for bilateral bargaining situations. For this purpose two bargaining environments were proposed which resemble commonly studied situations in bargaining literature. The first experiment examined the influence of risk-aversion on the bargaining share. To model risk-averse decision making within reinforcement learning we build agents on the basis of the architecture suggested in (Bellemare et al., 2017) that allows to learn the full value distribution. From this distribution a commonly used risk-metric can be derived, i.e. Conditional-Value at Risk. To study the influence of risk on the bargaining share, we matched agents with different degrees of risk-aversion and find that the bargaining share negatively correlates with an agents' risk-aversion. An interesting branch for further experiments would be to match agents with different risk metrics. Moreover, it would be interesting to analyze other possible influences on the bargaining outcome such as stubbornness. In this work we also ruled out the possibility of communication between agents which would also be an interesting aspect with respect to the bargaining behavior.

The second experiment aimed at analyzing the effect of information asymmetry on the general bargaining success. Information asymmetry here comes in the shape of quality uncertainty, i.e., a product may belong to different quality categories. Again two agents are involved in the bargaining process where agent 1 is the seller and agent 2 is the buyer of an item. The seller in this game holds information on the quality class of the current product whereas the buyer has no such information. From varying the product probabilities and the buyer's payoffs we find that quality uncertainty negatively affects the bargaining success.

All experimental results have to best of our knowledge not been demonstrated so far for multi-agent reinforcement learning.

REFERENCES

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in Economics*, pages 235–251. Elsevier.

Baarslag, T., Hendrikx, M. J., Hindriks, K. V., and Jonker, C. M. (2016). Learning about the opponent in automated bilateral negotiation: a comprehensive survey of opponent modeling techniques. *Autonomous Agents and Multi-Agent Systems*, 30(5):849–898.

Bellemare, M. G., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*.

Boutilier, C. (1996). Planning, learning and coordination in multiagent decision processes. In *Proceedings of the 6th conference on Theoretical aspects of rationality and knowledge*, pages 195–210. Morgan Kaufmann Publishers Inc.

Buşoniu, L., Babuška, R., and De Schutter, B. (2010). Multi-agent reinforcement learning: An overview. In *Innovations in multi-agent systems and applications-1*, pages 183–221. Springer.

Ellingsen, T. (1997). The evolution of bargaining behavior. *The Quarterly Journal of Economics*, 112(2):581–602.

Fang, F., Xin, Y., Yun, X., and Haitao, X. (2008). An opponent's negotiation behavior model to facilitate buyer-seller negotiations in supply chain management. In *2008 International Symposium on Electronic Commerce and Security*, pages 582–587. IEEE.

Fatima, S., Wooldridge, M., and Jennings, N. R. (2003). Comparing equilibria for game theoretic and evolutionary bargaining models. In *5th International Workshop on Agent-Mediated E-Commerce*, pages 70–77.

Fatima, S. S., Wooldridge, M., and Jennings, N. R. (2005). A comparative study of game theoretic and evolutionary models of bargaining for software agents. *Artificial Intelligence Review*, 23(2):187–205.

Garcıa, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

Gwak, J. and Sim, K. M. (2010). Bayesian learning based negotiation agents for supporting negotiation with incomplete information. In *World Congress on Engineering 2012. July 4-6, 2012. London, UK.*, volume 2188, pages 163–168. International Association of Engineers.

Haberland, V., Miles, S., and Luck, M. (2012). Adaptive negotiation for resource intensive tasks in grids. In *STAIRS*, pages 125–136.

Hayek, F. A. (1945). The use of knowledge in society. *The American economic review*, 35(4):519–530.

Kennan, J. and Wilson, R. (1993). Bargaining with private information. *Journal of Economic Literature*, 31(1):45–104.

Kisiala, J. (2015). Conditional value-at-risk: Theory and applications. *arXiv preprint arXiv:1511.00140*.

Konrad, K. A. and Morath, F. (2016). Bargaining with incomplete information: Evolutionary stability in finite populations. *Journal of Mathematical Economics*, 65:118–131.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473. International Foundation for Autonomous Agents and Multiagent Systems.

Lerer, A. and Peysakhovich, A. (2017). Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier.

Nash Jr, J. F. (1950). The bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 155–162.

Nguyen, T. T., Nguyen, N. D., and Nahavandi, S. (2018). Deep reinforcement learning for multi-agent systems: A review of challenges, solutions and applications. *arXiv preprint arXiv:1812.11794*.

Papaioannou, I. V., Roussaki, I. G., and Anagnostou, M. E. (2008). Neural networks against genetic algorithms for negotiating agent behaviour prediction. *Web Intelligence and Agent Systems: An International Journal*, 6(2):217–233.

Rand, D. G., Tarnita, C. E., Ohtsuki, H., and Nowak, M. A. (2013). Evolution of fairness in the one-shot anonymous ultimatum game. *Proceedings of the National Academy of Sciences*, 110(7):2581–2586.

Roth, A. E. (2012). *Axiomatic models of bargaining*, volume 170. Springer Science & Business Media.

Samuelson, W. (1984). Bargaining under asymmetric information. *Econometrica: Journal of the Econometric Society*, pages 995–1005.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tan, M. (1993). Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337.

Tuyls, K. and Nowé, A. (2005). Evolutionary game theory and multi-agent reinforcement learning. *The Knowledge Engineering Review*, 20(1):63–90.

Wang, W., Hao, J., Wang, Y., and Taylor, M. (2018). Towards cooperation in sequential prisoner's dilemmas: a deep multiagent reinforcement learning approach. *arXiv preprint arXiv:1803.00162*.