Risk Averse Reinforcement Learning for Mixed Multi-Agent Environments*

Extended Abstract

D. Sai Koti Reddy* Amrita Saha* Srikanth G. Tamilselvam
Priyanka Agrawal Pankaj Dayama
IBM Research
saikotireddy,amrsaha4,srikanth.tamilselvam,priyanka.agrawal,pankajdayama@in.ibm.com

ABSTRACT

Most real world applications of multi-agent systems, need to keep a balance between maximizing the rewards and minimizing the risks. In this work we consider a popular risk measure, variance of return (*VOR*), as a constraint in the agent's policy learning algorithm in the mixed cooperative and competitive environments. We present a multi-timescale actor critic method for risk sensitive Markov games where the risk is modeled as a *VOR* constraint. We also show that the risk-averse policies satisfy the desired risk constraint without compromising much on the overall reward for a popular task.

KEYWORDS

Risk-Sensitive Reinforcement Learning; Mixed Multi-Agent Environments; Actor-Critic Algorithms

ACM Reference Format:
D. Sai Koti Reddy* Amrita Saha* Srikanth G. Tamilselvam Priyanka Agrawal Pankaj Dayama. 2019. Risk Averse Reinforcement Learning for Mixed Multi-Agent Environments. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Reinforcement Learning (RL) has been an active area of research with practical implications to a diverse range of applications from gaming [14] to robotics [10] to finance [4]. Generally, RL algorithms [19] learn the optimal policy for an agent in a Markov Decision Process (MDP) framework to select actions that maximize the expected accumulated reward over time. However in many practical scenarios, agents may prefer a policy that provides lower expected reward but avoids uncertainty (i.e. actions with high but unpredictable rewards). The inherent uncertainty has been studied in detail by investigating the statistical properties of the return popularly known as *risk* in finance [13]. Some of the popular risk measures include variance-related measures [7, 18], Value-at-Risk (*VOR*) [6], Conditional Value-at-Risk (*CVaR*) [17] and percentile performance [12] etc. In this work, we consider the *VOR* risk measure.

The risk-sensitive objective in reinforcement learning is gaining traction in the recent times [5, 9]. Several variance-related risk measures and their corresponding algorithms to learn risk-sensitive policies were proposed in [16, 18, 20]. However, the prior works on risk-sensitive RL only deals with single-agent setting. To the best of our knowledge, ours is the first effort to look at risk-sensitive objective in multi-agent RL setting. Several practical applications like traffic control [21], coordination of autonomous vehicles [3] etc. implicitly involve interaction between multiple agents and one needs to account for the risk in such applications. We incorporate the variance-related risk as a constraint and model it as a constrained stochastic game. We consider mixed cooperative and competitive multi-agent setting and develop actor-critic algorithms under the the centralized training with decentralized execution framework [8, 11].

*equal contribution by the first two authors.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

2 MODEL

The multi-agent environments are modeled as a stochastic (Markov) game G which is defined as a tuple $(S, A, \mathcal{T}, r, Z, O, N, \gamma)$, where S denotes the states in the environment. The agents $\{1, ..., N\}$ upon taking actions from the set $\{A_1, ..., A_k\}$, move to the next state governed by environment's state transition function $\mathcal{T} : S \times A_1 \times ... \times A_k \rightarrow S$. At each time step, the agents draw partial observation $z \in Z$ based on observation function $O(s, n) : S \times A \rightarrow Z$. Each agent also maintains an action-observation sequence $H = (Z \times A)^*$ on which it conditions a stochastic policy $\pi^\theta : H \times A \rightarrow [0, 1]$ to take an action. This results in a cost based on the cost function $r_i(s, a) : S \times A_n \rightarrow \mathbb{R}$ for an agent i and discount factor $\gamma \in [0, 1)$.

The goal for the agent i is to minimize the expected discounted cost: $R_i = \Sigma_{t=0}^{T} \gamma^t r_i^t$. Here, r_i is the cost function for agent i. The discounted return of state s_i and state-action pair (s_i, a_i) for agent i is defined as $R_i(s_i) = \Sigma_{t=0}^{T} \gamma^t r_i^t(s^t, a^t)|s_0 = s_i, \pi^{\theta_i}$ and $R_i(s_i, a_i) = \Sigma_{t=0}^{T} \gamma^t r_i^t(s^t, a^t)|s_0 = s_i, a_0 = a_i, \pi^{\theta_i}$ respectively. For an agent i, the expected value of the return for state s and state-action pair (s, a) are known as value function $V_i^\theta(s) = \mathbb{E}[R_i(s)]$ and action-value function $Q_i^\theta(s, a) = \mathbb{E}[R_i(s, a)]$ respectively, for a policy π^θ.

3 VOR RISK CONSTRAINED OBJECTIVE

We present *VOR* [18] as a variance-related risk measure for multi-agent mixed cooperative and competitive setting. VOR is a measure of variability in the reward sequence for a policy π^θ, defined as:

$$\Lambda^\theta(s) = \mathbb{E}[R^\theta(s)^2] - V^\theta(s)^2, \tag{1}$$

We define square reward function as follows: $U^\theta(s) := \mathbb{E}[R^\theta(s)^2]$. In risk neural settings, the goal of agent i is to find the optimal policy parameter by solving the following equation:

$$\theta_i^* = \arg\min_{\theta_i} J(\theta_i), \quad \text{where} \quad J(\theta_i) = \mathbb{E}[R^{\theta_i}] \tag{2}$$

Algorithm 1 Risk Constrained MADDPG

for episode = 1 to M **do**
  Input ← Initial State s and Initialize Replay Buffer $\mathcal{D}$
  for n = 1 to *max episode length* **do**
    for each agent i, select a action $a_i = \mu_{\theta_i}(o_i) + \mathcal{N}_t$ { where $\mathcal{N}_t$ is a random process and o_i is partial state observation of agent i }
    Execute actions $a = (a_1, \ldots, a_N)$ from state s and observe reward r by going to new state s'
    Store (s, a, r, s') in replay buffer $\mathcal{D}$ and $s \leftarrow s'$
    for agent i = 1 to N **do**
      Sample a random minibatch of S samples (s^j, a^j, r^j, s'^j) from $\mathcal{D}$
      Estimate $\Lambda^{\theta_i}(s^j)$ and set $y^j = r_i^j + \gamma Q_i^{\mu'}(s'^j, a'_1, \ldots, a'_N)|_{a'_k=\mu'_k(o_k^j)} + \lambda_i(\Lambda^{\theta_i}(s^j) - \delta_i)$ { μ' is the deterministic policy }
      Update critic by minimizing the loss $\mathcal{L}(\theta_i) = \frac{1}{S}\Sigma_j(y^j - Q_i^{\mu}(s^j, a_1^j, \ldots, a_N^j))^2$ { Q_i^{μ} is central action-value function of agent i }
      Gradient descent for actor's policy parameter on faster timescale: $\nabla_{\theta_i} L \approx \frac{1}{S}\Sigma_j \nabla_{\theta_i}\mu_i(o_i^j)\nabla_{a_i} Q_i^{\mu}(s^j, a_1^j, \ldots, a_i, \ldots, a_N^j)|_{a_i=\mu_i(o_i^j)}$
      $\lambda_i^{n+1} = \max(0, \lambda_i^n + b^n(\Lambda^{\theta_i}(s^i) - \delta_i)$ where b^n is the step size { Gradient ascent for Lagrange multiplier on slower timescale }
    end for
    Update target network parameters for each agent i: $\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$
  end for
end for=0

However, in our paper the goal is to find the risk averse policy by including VOR risk measure as a constraint. VOR risk-sensitive objective function for an agent i is defined as follows:

$$\min_{\theta_i} J(\theta_i) \text{ s.t. } \Lambda^{\theta_i}(s_i^0) \leq \delta_i \tag{3}$$

We employ the Lagrangian relaxation procedure [1] which converts the constrained optimization problem (3) into the following unconstrained optimization problem:

$$\max_{\lambda_i \geq 0} \min_{\theta_i} \left(L(\theta_i, \lambda_i) := J(\theta_i) + \lambda_i \left(\Lambda^{\theta_i}(s_i^0) - \delta_i \right) \right) \tag{4}$$

where λ_i is the Lagrange multiplier. The objective for an agent i is to find the local saddle point $(\theta_i^*, \lambda_i^*)$ of Lagrangian (4) which satisfies (5). From the saddle point theorem, the θ_i^* is local optimal policy parameter for VOR-constrained optimization problem (3).

$$L(\theta_i, \lambda_i) \geq L(\theta_i^*, \lambda_i^*) \geq L(\theta_i^*, \lambda_i) \tag{5}$$

4 PROPOSED ALGORITHM

We propose RC-MADDPG actor-critic algorithm for finding solution $(\theta_i^*, \lambda_i^*)$. The RC-MADDPG algorithm uses multi-timescale approach [2] along with the centralized training with decentralized execution framework similar to [8, 11]. Details of our algorithm are given in Algorithm 1.

5 EXPERIMENTS

We considered well-studied mixed multi-agent environment **Keep Away** [11][15], where the goal of good agents is to reach the landmark while adversaries are pushing them away.

Training Details: For the above task, we vary both the number of good agents and adversaries upto 2, resulting in different multi-agent scenarios with maximum of 4 agents. The main parameters are the policy parameter θ, the Lagrange multiplier λ, which are updated in a multi-timescale based updation strategy.

Estimation of risk-criteria in unconstrained case: For the Keep-Away task, we train a multi-agent RL model in an unconstrained setting for around 60K episodes with MADDPG algorithm [11] and used this converged policy for the next 1000 episodes to estimate the reward and VOR in the unconstrained setting. Next we train our proposed risk-constrained algorithm and estimate the reward and VOR for the constrained case. We used half of the estimated VOR for converged unconstrained policy as a constraint VOR-δ in our VOR risk-constrained objective for each agent.

Results: In each of the scenarios, we compare the proposed RC-MADDPG algorithm for VOR risk averse policy with risk neutral policy obtained by [11]. Each experiment is repeated 30 times and their average is reported in Table 1. Comparative analysis of the results of the constrained model with the unconstrained one results show that it is indeed possible to generate constraint satisfying policies that can achieve similar reward as their respective unconstrained counterpart.

N,M	Unconstrained		VOR constrained		
	VOR	Reward	VOR-δ	VOR	Reward
1, 1	0.653	-9.293	0.327	**0.276**	-11.036
1, 2	1.2	-19.187	0.6	**0.532**	-20.542
2, 1	0.749	-12.519	0.375	**0.246**	-15.689
2, 2	1.385	-26.587	0.693	**0.477**	-30.403

Table 1: Comparison of the VOR constrained policy with the unconstrained one. N and M are the number of adversaries and good agents. VOR-δ is the upper limit of VOR in the constrained case. Bold numbers indicate that the constraint has been satisfied.

6 CONCLUSION

We considered the problem of finding the risk averse polices for mixed multi-agent cooperative-competitive environments. We incorporated VOR risk measure as a constraint and proposed RC-MADDPG algorithm that adopts "centralized training of decentralized policies" framework and multi-timescale approach. We empirically demonstrated how the constrained policy can be risk-averse and yet achieve similar rewards as the unconstrained one, on the Keep Away task.

REFERENCES

[1] Dimitri P Bertsekas. 1999. *Nonlinear programming*. Athena scientific Belmont.

[2] Vivek S Borkar. 2009. *Stochastic approximation: a dynamical systems viewpoint*. Vol. 48. Springer.

[3] Yongcan Cao, Wenwu Yu, Wei Ren, and Guanrong Chen. 2013. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial informatics* 9, 1 (2013), 427–438.

[4] James J Choi, David Laibson, Brigitte C Madrian, and Andrew Metrick. 2009. Reinforcement learning and savings behavior. *The Journal of finance* 64, 6 (2009), 2515–2534.

[5] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. 2018. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research* 18, 167 (2018), 1–51.

[6] Darrell Duffie and Jun Pan. 1997. An overview of value at risk. *Journal of derivatives* 4, 3 (1997), 7–49.

[7] Jerzy A Filar, Lodewijk CM Kallenberg, and Huey-Miin Lee. 1989. Variance-penalized Markov decision processes. *Mathematics of Operations Research* 14, 1 (1989), 147–161.

[8] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2017. Counterfactual multi-agent policy gradients. *arXiv preprint arXiv:1705.08926* (2017).

[9] Michael Fu et al. 2018. Risk-Sensitive Reinforcement Learning: A Constrained Optimization Viewpoint. *arXiv preprint arXiv:1810.09126* (2018).

[10] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. 2016. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research* 17, 1 (2016), 1334–1373.

[11] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*. 6379–6390.

[12] John B Mander, Rajesh P Dhakal, Naoto Mashiko, and Kevin M Solberg. 2007. Incremental dynamic analysis applied to seismic financial risk assessment of bridges. *Engineering structures* 29, 10 (2007), 2662–2672.

[13] Harry Markowitz. 1952. Portfolio selection. *The journal of finance* 7, 1 (1952), 77–91.

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.

[15] Igor Mordatch and Pieter Abbeel. 2017. Emergence of Grounded Compositional Language in Multi-Agent Populations. *arXiv preprint arXiv:1703.04908* (2017).

[16] LA Prashanth and Mohammad Ghavamzadeh. 2013. Actor-critic algorithms for risk-sensitive MDPs. In *Advances in neural information processing systems*. 252–260.

[17] R Tyrrell Rockafellar, Stanislav Uryasev, et al. 2000. Optimization of conditional value-at-risk. *Journal of risk* 2 (2000), 21–42.

[18] Matthew J Sobel. 1982. The variance of discounted Markov decision processes. *Journal of Applied Probability* 19, 4 (1982), 794–802.

[19] Richard S Sutton and Andrew G Barto. 1998. *Introduction to reinforcement learning*. Vol. 135. MIT press Cambridge.

[20] Aviv Tamar, Dotan Di Castro, and Shie Mannor. 2012. Policy gradients with variance related risk criteria. In *Proceedings of the twenty-ninth international conference on machine learning*. 387–396.

[21] MA Wiering. 2000. Multi-agent reinforcement learning for traffic light control. In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*. 1151–1158.