AAAI-2021论文整理

> 本篇文档主要是整理2021年中在被AAAI所接受的相关论文(**POMDP、Multi-Agent、RL**)，每篇文章包含**标题**、**摘要**和**论文主页链接**。
>
> 每个方向的论文被分别整理到相关方向下，若同一篇论文属于多个方向，则优先放入**Multi-Agent**，其次为**POMDP**。其中第一作者为国内学术机构的文章标题被高亮显示。
>
> **AAAI2021论文接受列表PDF**
>
> **AAAI2021论文主页**

POMDP

POMDP部分收录了：

1. AAAI Technical Track on Machine Learning I-V中的1篇文章(1)。
2. AAAI Technical Track on Planning, Routing, and Scheduling中的5篇文章(2-8)。
3. AAAI Technical Track on Reasoning under Uncertainty中的2篇文章(9-10)。

1.Deep Recurrent Belief Propagation Network for POMDPs

Abstract: In many real-world sequential decision-making tasks, especially in continuous control like robotic control, it is rare that the observations are perfect, that is, the sensory data could be incomplete, noisy or even dynamically polluted due to the unexpected malfunctions or intrinsic low quality of the sensors. Previous methods handle these issues in the framework of POMDPs and are either deterministic by feature memorization or stochastic by belief inference. In this paper, we present a new method that lies somewhere in the middle of the spectrum of research methodology identified above and combines the strength of both approaches. In particular, the proposed method, named Deep Recurrent Belief Propagation Network (DRBPN), takes a hybrid style belief updating procedure – an RNN-type feature extraction step followed by an analytical belief inference, significantly reducing the computational cost while faithfully capturing the complex dynamics and maintaining the necessary uncertainty for generalization. The effectiveness of the proposed method is verified on a collection of benchmark tasks, showing that our approach outperforms several state-of-the-art methods under various challenging scenarios.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17227

2.Minimax Regret Optimisation for Robust Planning in Uncertain Markov Decision Processes

Abstract: The parameters for a Markov Decision Process (MDP) often cannot be specified exactly. Uncertain MDPs (UMDPs) capture this model ambiguity by defining sets which the parameters belong to. Minimax regret has been proposed as an objective for planning in UMDPs to find robust policies which are not overly conservative. In this work, we focus on planning for Stochastic Shortest Path (SSP) UMDPs with uncertain cost and transition functions. We introduce a Bellman equation to compute the regret for a policy. We propose a dynamic programming algorithm that utilises the regret Bellman equation, and show that it optimises minimax regret exactly for UMDPs with independent uncertainties. For coupled uncertainties, we extend our approach to use options to enable a trade off between computation and solution quality. We

evaluate our approach on both synthetic and real-world domains, showing that it significantly outperforms existing baselines.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17417

3.Dynamic Automaton-Guided Reward Shaping for Monte Carlo Tree Search

Abstract: Reinforcement learning and planning have been revolutionized in recent years, due in part to the mass adoption of deep convolutional neural networks and the resurgence of powerful methods to refine decision-making policies. However, the problem of sparse reward signals and their representation remains pervasive in many domains. While various rewardshaping mechanisms and imitation learning approaches have been proposed to mitigate this problem, the use of humanaided artificial rewards introduces human error, sub-optimal behavior, and a greater propensity for reward hacking. In this paper, we mitigate this by representing objectives as automata in order to define novel reward shaping functions over this structured representation. In doing so, we address the sparse rewards problem within a novel implementation of Monte Carlo Tree Search (MCTS) by proposing a reward shaping function which is updated dynamically to capture statistics on the utility of each automaton transition as it pertains to satisfying the goal of the agent. We further demonstrate that such automaton-guided reward shaping can be utilized to facilitate transfer learning between different environments when the objective is the same.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17427

4.Constrained Risk-Averse Markov Decision Processes

Abstract: We consider the problem of designing policies for Markov decision processes (MDPs) with dynamic coherent risk objectives and constraints. We begin by formulating the problem in a Lagrangian framework. Under the assumption that the risk objectives and constraints can be represented by a Markov risk transition mapping, we propose an optimization-based method to synthesize Markovian policies that lower-bound the constrained risk-averse problem. We demonstrate that the formulated optimization problems are in the form of difference convex programs (DCPs) and can be solved by the disciplined convex-concave programming (DCCP) framework. We show that these results generalize linear programs for constrained MDPs with total discounted expected costs and constraints. Finally, we illustrate the effectiveness of the proposed method with numerical experiments on a rover navigation problem involving conditional-value-at-risk (CVaR) and entropic-value-at-risk (EVaR) coherent risk measures.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17393

5.Successor Feature Sets: Generalizing Successor Representations Across Policies

Abstract: Successor-style representations have many advantages for reinforcement learning: for example, they can help an agent generalize from past experience to new goals, and they have been proposed as explanations of behavioral and neural data from human and animal learners. They also form a natural bridge between model-based and model-free RL methods: like the former they make predictions about future experiences, and like the latter they allow efficient prediction of total discounted rewards. However, successor-style representations are not optimized to generalize across policies: typically, we maintain a limited-length list of policies, and share information among them by representation learning or GPI. Successor-style representations also typically make no provision for gathering information or reasoning about latent variables. To address these limitations, we bring together ideas from predictive state representations, belief space value iteration, successor features, and convex analysis: we develop a new, general successor-style representation, together with a Bellman equation that connects multiple sources of information within this representation, including different latent states, policies, and reward functions. The new representation is highly expressive: for example, it lets us efficiently read off an optimal policy for a new reward function, or a policy that imitates a new demonstration. For this paper, we focus on exact computation of the new representation in small, known environments, since even this restricted setting offers plenty of interesting questions. Our implementation does not scale to large, unknown environments --- nor would we expect it to, since it generalizes POMDP value iteration, which is difficult to scale. However, we believe that future work will allow us to extend our ideas to approximate reasoning in large, unknown environments. We conduct experiments to explore which of the potential barriers to scaling are most pressing.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17399

6.Robust Finite-State Controllers for Uncertain POMDPs

Abstract: Uncertain partially observable Markov decision processes (uPOMDPs) allow the probabilistic transition and observation functions of standard POMDPs to belong to a so-called uncertainty set. Such uncertainty, referred to as epistemic uncertainty, captures uncountable sets of probability distributions caused by, for instance, a lack of data available. We develop an algorithm to compute finite-memory policies for uPOMDPs that robustly satisfy specifications against any admissible distribution. In general, computing such policies is theoretically and practically intractable. We provide an efficient solution to this problem in four steps. (1) We state the underlying problem as a nonconvex optimization problem with infinitely many constraints. (2) A dedicated dualization scheme yields a dual problem that is still nonconvex but has finitely many constraints. (3) We linearize this dual problem and (4) solve the resulting finite linear program to obtain locally optimal solutions to the original problem. The resulting problem formulation is exponentially smaller than those resulting from existing methods. We demonstrate the applicability of our algorithm using large instances of an aircraft collision-avoidance scenario and a novel spacecraft motion planning case study.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17401

7.Bayesian Optimized Monte Carlo Planning

Abstract: Online solvers for partially observable Markov decision processes have difficulty scaling to problems with large action spaces. Monte Carlo tree search with progressive widening attempts to improve scaling by sampling from the action space to construct a policy search tree. The performance of progressive widening search is dependent upon the action sampling policy, often requiring problem-specific samplers. In this work, we present a general method for efficient action sampling based on Bayesian optimization. The proposed method uses a Gaussian process to model a belief over the action-value function and selects the action that will maximize the expected improvement in the optimal action value. We implement the proposed approach in a new online tree search algorithm called Bayesian Optimized Monte Carlo Planning (BOMCP). Several experiments show that BOMCP is better able to scale to large action space POMDPs than existing state-of-the-art tree search solvers.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17411

8.Improved POMDP Tree Search Planning with Prioritized Action Branching

Abstract: Online solvers for partially observable Markov decision processes have difficulty scaling to problems with large action spaces. This paper proposes a method called PA-POMCPOW to sample a subset of the action space that provides varying mixtures of exploitation and exploration for inclusion in a search tree. The proposed method first evaluates the action space according to a score function that is a linear combination of expected reward and expected information gain. The actions with the highest score are then added to the search tree during tree expansion. Experiments show that PA-POMCPOW is able to outperform existing state-of-the-art solvers on problems with large discrete action spaces.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17412

9.Scalable First-Order Methods for Robust MDPs

Abstract: Robust Markov Decision Processes (MDPs) are a powerful framework for modeling sequential decision making problems with model uncertainty. This paper proposes the first first-order framework for solving robust MDPs. Our algorithm interleaves primal-dual first-order updates with approximate Value Iteration updates. By carefully controlling the tradeoff between the accuracy and cost of Value Iteration updates, we achieve an ergodic convergence rate that is significantly better than classical Value Iteration algorithms in terms of the number of states S and the number of actions A on ellipsoidal and Kullback-Leibler s-rectangular uncertainty sets. In numerical experiments on ellipsoidal uncertainty sets we show that our algorithm is significantly more scalable than state-of-the-art approaches. Our framework is also the first one to solve robust MDPs with s-rectangular KL uncertainty sets.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17435

10.Robust Contextual Bandits via Bootstrapping

Abstract: Upper confidence bound (UCB) based contextual bandit algorithms require one to know the tail property of the reward distribution. Unfortunately, such tail property is usually unknown or difficult to specify in real-world applications. Using a tail property heavier than the ground truth leads to a slow learning speed of the contextual bandit algorithm, while using a lighter one may cause the algorithm to diverge. To address this fundamental problem, we develop an estimator (evaluated from historical rewards) for the contextual bandit UCB based on the multiplier bootstrapping technique. We first establish sufficient conditions under which our estimator converges asymptotically to the ground truth of contextual bandit UCB. We further derive a second order correction for our estimator so as to obtain its confidence level with a finite number of rounds. To demonstrate the versatility of the estimator, we apply it to design a BootLinUCB algorithm for the contextual bandit. We prove that the BootLinUCB has a sub-linear regret upper bound and also conduct extensive experiments to validate its superior performance.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17446

Multi-Agent

Multi-Agent部分收录了:

1. AAAI Technical Track on Machine Learning I-V中的4篇文章(1-4)
2. AAAI Technical Track on Multiagent Systems中的20篇文章(5-24)

3. AAAI Technical Track on Planning, Routing, and Scheduling中的1篇文章(25)

1.Decentralized Multi-Agent Linear Bandits with Safety Constraints

Abstract: We study decentralized stochastic linear bandits, where a network of N agents acts cooperatively to efficiently solve a linear bandit-optimization problem over a d-dimensional space. For this problem, we propose DLUCB: a fully decentralized algorithm that minimizes the cumulative regret over the entire network. At each round of the algorithm each agent chooses its actions following an upper confidence bound (UCB) strategy and agents share information with their immediate neighbors through a carefully designed consensus procedure that repeats over cycles. Our analysis adjusts the duration of these communication cycles ensuring near-optimal regret performance O(d \log{NT}\sqrt{NT}) at a communication rate of O(dN^2) per round. The structure of the network affects the regret performance via a small additive term – coined the regret of delay – that depends on the spectral gap of the underlying graph. Notably, our results apply to arbitrary network topologies without a requirement for a dedicated agent acting as a server. In consideration of situations with high communication cost, we propose RC-DLUCB: a modification of DLUCB with rare communication among agents. The new algorithm trades off regret performance for a significantly reduced total communication cost of O(d^{3N}5/2) over all T rounds. Finally, we show that our ideas extend naturally to the emerging, albeit more challenging, setting of safe bandits. For the recently studied problem of linear bandits with unknown linear safety constraints, we propose the first safe decentralized algorithm. Our study contributes towards applying bandit techniques in safety-critical distributed systems that repeatedly deal with unknown stochastic environments. We present numerical simulations for various network topologies that corroborate our theoretical findings.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16820

2.Solving Common-Payoff Games with Approximate Policy Iteration

Abstract: For artificially intelligent learning systems to have widespread applicability in real-world settings, it is important that they be able to operate decentrally. Unfortunately, decentralized control is difficult---computing even an epsilon-optimal joint policy is a NEXP complete problem. Nevertheless, a recently rediscovered insight---that a team of agents can coordinate via common knowledge---has given rise to algorithms capable of finding optimal joint policies in small common-payoff games. The Bayesian action decoder (BAD) leverages this insight and deep reinforcement learning to scale to games as large as two-player Hanabi. However, the approximations it uses to do so prevent it from discovering optimal joint policies even in games small enough to brute force optimal solutions. This work proposes CAPI, a novel algorithm which, like BAD, combines common knowledge with deep reinforcement learning. However, unlike BAD, CAPI prioritizes the propensity to discover optimal joint policies over scalability. While this choice precludes CAPI from scaling to games as large as Hanabi, empirical results demonstrate that, on the games to which CAPI does scale, it is capable of discovering optimal joint policies even when other modern multi-agent reinforcement learning algorithms are unable to do so.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17166

3.Reinforcement Learning Based Multi-Agent Resilient Control: From Deep Neural Networks to an Adaptive Law

Abstract: Recent advances in Multi-agent Reinforcement Learning (MARL) have made it possible to implement various tasks in cooperative as well as competitive scenarios through trial and error, and deep neural networks. These successes motivate us to bring the mechanism of MARL into the Multi-agent Resilient Consensus (MARC) problem that studies the consensus problem in a network of agents with faulty ones. Relying on the natural characteristics of the system goal, the key component in MARL, reward function, can thus be directly constructed via the relative distance among agents. Firstly, we apply Deep Deterministic Policy Gradient (DDPG) on each single agent to train and learn adjacent weights of neighboring agents in a distributed manner, that we call Distributed-DDPG (D-DDPG), so as to minimize the weights from suspicious agents and eliminate the corresponding influences. Secondly, to get rid of neural networks and their time-consuming training process, a Q-learning based algorithm, called Q-consensus, is further presented by building a proper reward function and a credibility function for each pair of neighboring agents so that the adjacent weights can update in an adaptive way. The experimental results indicate that both algorithms perform well with appearance of constant and/or random faulty agents, yet the Q-consensus algorithm outperforms the faulty ones running D-DDPG. Compared to the traditional resilient consensus strategies, e.g., Weighted-Mean-Subsequence-Reduced (W-MSR) or trustworthiness analysis, the proposed Q-consensus algorithm has greatly relaxed the topology requirements, as well as reduced the storage and computation loads. Finally, a smart-car hardware platform consisting of six vehicles is used to verify the effectiveness of the Q-consensus algorithm by achieving resilient velocity synchronization.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16945

4.Decentralized Policy Gradient Descent Ascent for Safe Multi-Agent Reinforcement Learning

Abstract: This paper deals with distributed reinforcement learning problems with safety constraints. In particular, we consider that a team of agents cooperate in a shared environment, where each agent has its individual reward function and safety constraints that involve all agents' joint actions. As such, the agents aim to maximize the team-average long-term return, subject to all the safety constraints. More intriguingly, no central controller is assumed to coordinate the agents, and both the rewards and constraints are only known to each agent locally/privately. Instead, the agents are connected by a peer-to-peer communication network to share information with their neighbors. In this work, we first formulate this problem as a distributed constrained Markov decision process (D-CMDP) with networked agents. Then, we propose a decentralized policy gradient (PG) method, Safe Dec-PG, to perform policy optimization based on this D-CMDP model over a network. Convergence guarantees, together with numerical results, showcase the superiority of the proposed algorithm. To the best of our knowledge, this is the first decentralized PG algorithm that accounts for the coupled safety constraints with a quantifiable convergence rate in multi-agent reinforcement learning. Finally, we emphasize that our algorithm is also novel in solving a class of decentralized stochastic nonconvex-concave minimax optimization problems, where both the algorithm design and corresponding theoretical analysis are of independent interest.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17062

5.Improving Continuous-time Conflict Based Search

Abstract: Conflict-Based Search (CBS) is a powerful algorithmic framework for optimally solving classical multi-agent path finding (MAPF) problems, where time is discretized into the time steps. Continuous-time CBS (CCBS) is a recently proposed version of CBS that guarantees optimal solutions without the need to discretize time. However, the scalability of CCBS is limited because it does not include any known improvements of CBS. In this paper, we begin to close this gap and explore how to adapt successful CBS improvements, namely, prioritizing conflicts (PC), disjoint splitting (DS), and high-level heuristics, to the continuous time setting of CCBS. These adaptions are not trivial, and require careful handling of different types of constraints, applying a generalized version of the Safe interval path planning (SIPP) algorithm, and extending the notion of cardinal conflicts. We evaluate the effect of the suggested enhancements by running experiments both on general graphs and 2^k-neighborhood grids. CCBS with these improvements significantly outperforms vanilla CCBS, solving problems with almost twice as many agents in some cases and pushing the limits of multi-agent path finding in continuous-time domains.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17338

6.Inference-Based Deterministic Messaging For Multi-Agent Communication

Abstract: Communication is essential for coordination among humans and animals. Therefore, with the introduction of intelligent agents into the world, agent-to-agent and agent-to-human communication becomes necessary. In this paper, we first study learning in matrix-based signaling games to empirically show that decentralized methods can converge to a suboptimal policy. We then propose a modification to the messaging policy, in which the sender deterministically chooses the best message that helps the receiver to infer the sender's observation. Using this modification, we see, empirically, that the agents converge to the optimal policy in nearly all the runs. We then apply this method to a partially observable gridworld environment which requires cooperation between two agents and show that, with appropriate approximation methods, the proposed sender modification can enhance existing decentralized training methods for more complex domains as well.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17339

7.Scalable and Safe Multi-Agent Motion Planning with Nonlinear Dynamics and Bounded Disturbances

Abstract: We present a scalable and effective multi-agent safe motion planner that enables a group of agents to move to their desired locations while avoiding collisions with obstacles and other agents, with the presence of rich obstacles, high-dimensional, nonlinear, nonholonomic dynamics, actuation limits, and disturbances. We address this problem by finding a piecewise linear path for each agent such that the actual trajectories following these paths are guaranteed to satisfy the reach-and-avoid requirement. We show that the spatial tracking error of the actual trajectories of the controlled agents can be pre-computed for any qualified path that considers the minimum duration of each path segment due to actuation limits. Using these bounds, we find a collision-free path for each agent by solving Mixed Integer-Linear Programs and coordinate agents by using the priority-based search. We demonstrate our method by benchmarking in 2D and 3D scenarios with ground vehicles and quadrotors, respectively, and show improvements over the solving time and the solution quality compared to two state-of-the-art multi-agent motion planners.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17340

8.Learning to Resolve Conflicts for Multi-Agent Path Finding with Conflict-Based Search

Abstract: Conflict-Based Search (CBS) is a state-of-the-art algorithm for multi-agent path finding. On the high level, CBS repeatedly detects conflicts and resolves one of them by splitting the current problem into two subproblems. Previous work chooses the conflict to resolve by categorizing conflicts into three classes and always picking one from the highest-priority class. In this work, we propose an oracle for conflict selection that results in smaller search tree sizes than the one used in previous work. However, the computation of the oracle is slow. Thus, we propose a machine-learning (ML) framework for conflict selection that observes the decisions made by the oracle and learns a conflict-selection strategy represented by a linear ranking function that imitates the oracle's decisions accurately and quickly. Experiments on benchmark maps indicate that our approach, ML-guided CBS, significantly improves the success rates, search tree sizes and runtimes of the current state-of-the-art CBS solver.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17341

9.The Influence of Memory in Multi-Agent Consensus

Abstract: Multi-agent consensus problems can often be seen as a sequence of autonomous and independent local choices between a finite set of decision options, with each local choice undertaken simultaneously, and with a shared goal of achieving a global consensus state. Being able to estimate probabilities for the different outcomes and to predict how long it takes for a consensus to be formed, if ever, are core issues for such protocols. Little attention has been given to protocols in which agents can remember past or outdated states. In this paper, we propose a framework to study what we call `memory consensus protocol'. We show that the employment of memory allows such processes to always converge, as well as, in some scenarios, such as cycles, converge faster. We provide a theoretical analysis of the probability of each option eventually winning such processes based on the initial opinions expressed by agents. Further, we perform experiments to investigate network topologies in which agents benefit from memory on the expected time needed for consensus.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17342

10.Exploration-Exploitation in Multi-Agent Learning: Catastrophe Theory Meets Game Theory

Abstract: Exploration-exploitation is a powerful and practical tool in multi-agent learning (MAL), however, its effects are far from understood. To make progress in this direction, we study a smooth analogue of Q-learning. We start by showing that our learning model has strong theoretical justification as an optimal model for studying exploration-exploitation. Specifically, we prove that smooth Q-learning has bounded regret in arbitrary games for a cost model that explicitly captures the balance between game and exploration costs and that it always converges to the set of quantal-response equilibria (QRE), the standard solution concept for games under bounded rationality, in weighted potential games with heterogeneous learning agents. In our main task, we then turn to measure the effect of exploration in collective system performance. We characterize the geometry of the QRE surface in low-dimensional MAL systems and link our findings with catastrophe (bifurcation) theory. In particular, as the exploration hyperparameter evolves over-time, the system undergoes phase transitions where the number and stability of equilibria can change radically given an infinitesimal change to the exploration parameter. Based on this, we provide a formal theoretical treatment of how tuning the exploration parameter can provably lead to equilibrium selection with both positive as well as negative (and potentially unbounded) effects to system performance.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17343

11.Lifelong Multi-Agent Path Finding in Large-Scale Warehouses

Abstract: Multi-Agent Path Finding (MAPF) is the problem of moving a team of agents to their goal locations without collisions. In this paper, we study the lifelong variant of MAPF, where agents are constantly engaged with new goal locations, such as in large-scale automated warehouses. We propose a new framework Rolling-Horizon Collision Resolution (RHCR) for solving lifelong MAPF by decomposing the problem into a sequence of Windowed MAPF instances, where a Windowed MAPF solver resolves collisions among the paths of the agents only within a bounded time horizon and ignores collisions beyond it. RHCR is particularly well suited to generating pliable plans that adapt to continually arriving new goal locations. We empirically evaluate RHCR with a variety of MAPF solvers and show that it can produce high-quality solutions for up to 1,000 agents (= 38.9% of the empty cells on the map) for simulated warehouse instances, significantly outperforming existing work.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17344

12.Dec-SGTS: Decentralized Sub-Goal Tree Search for Multi-Agent Coordination

Abstract: Multi-agent coordination tends to benefit from efficient communication, where cooperation often happens based on exchanging information about what the agents intend to do, i.e. intention sharing. It becomes a key problem to model the intention by some proper abstraction. Currently, it is either too coarse such as final goals or too fined as primitive steps, which is inefficient due to the lack of modularity and semantics. In this paper, we design a novel multi-agent coordination protocol based on subgoal intentions, defined as the probability distribution over feasible subgoal sequences. The subgoal intentions encode macro-action behaviors with modularity so as to facilitate joint decision making at higher abstraction. Built over the proposed protocol, we present Dec-SGTS (Decentralized Sub-Goal Tree Search) to solve decentralized online multi-agent planning hierarchically and efficiently. Each agent runs Dec-SGTS asynchronously by iteratively performing three phases including local sub-goal tree search, local subgoal intention update and global subgoal intention sharing. We conduct the experiments on courier dispatching problem, and the results show that Dec-SGTS achieves much better reward while enjoying a significant reduction of planning time and communication cost compared with Dec-MCTS (Decentralized Monte Carlo Tree Search).

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17345

13.Expected Value of Communication for Planning in Ad Hoc Teamwork

Abstract: A desirable goal for autonomous agents is to be able to coordinate on the fly with previously unknown teammates. Known as “ad hoc teamwork”, enabling such a capability has been receiving increasing attention in the research community. One of the central challenges in ad hoc teamwork is quickly recognizing the current plans of other agents and planning accordingly. In this paper, we focus on the scenario in which teammates can communicate with one another, but only at a cost. Thus, they must carefully balance plan recognition based on observations vs. that based on communication. This paper proposes a new metric for evaluating how similar are two policies that a teammate may be following - the Expected Divergence Point (EDP). We then present a novel planning algorithm for ad hoc teamwork, determining which query to ask and planning accordingly. We demonstrate the effectiveness of this algorithm in a range of increasingly general communication in ad hoc teamwork problems.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17346

14.Time-Independent Planning for Multiple Moving Agents

Abstract: Typical Multi-agent Path Finding (MAPF) solvers assume that agents move synchronously, thus neglecting the reality gap in timing assumptions, e.g., delays caused by an imperfect execution of asynchronous moves. So far, two policies enforce a robust execution of MAPF plans taken as input: either by forcing agents to synchronize or by executing plans while preserving temporal dependencies. This paper proposes an alternative approach, called time-independent planning, which is both online and distributed. We represent reality as a transition system that changes configurations according to atomic actions of agents, and use it to generate a time-independent schedule. Empirical results in a simulated environment with stochastic delays of agents' moves support the validity of our proposal.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17347

15.Resilient Multi-Agent Reinforcement Learning with Adversarial Value Decomposition

Abstract: We focus on resilience in cooperative multi-agent systems, where agents can change their behavior due to udpates or failures of hardware and software components. Current state-of-the-art approaches to cooperative multi-agent reinforcement learning (MARL) have either focused on idealized settings without any changes or on very specialized scenarios, where the number of changing agents is fixed, e.g., in extreme cases with only one productive agent. Therefore, we propose Resilient Adversarial value Decomposition with Antagonist-Ratios (RADAR). RADAR offers a value decomposition scheme to train competing teams of varying size for improved resilience against arbitrary agent changes. We evaluate RADAR in two cooperative multi-agent domains and show that RADAR achieves better worst case performance w.r.t. arbitrary agent changes than state-of-the-art MARL.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17348

16.Anytime Heuristic and Monte Carlo Methods for Large-Scale Simultaneous Coalition Structure Generation and Assignment

Abstract: Optimal simultaneous coalition structure generation and assignment is computationally hard. The state-of-the-art can only compute solutions to problems with severely limited input sizes, and no effective approximation algorithms that are guaranteed to yield high-quality solutions are expected to exist. Real-world optimization problems, however, are often characterized by large-scale inputs and the need for generating feasible solutions of high quality in limited time. In light of this, and to make it possible to generate better feasible solutions for difficult large-scale problems efficiently, we present and benchmark several different anytime algorithms that use general-purpose heuristics and Monte Carlo techniques to guide search. We evaluate our methods using synthetic problem sets of varying distribution and complexity. Our results show that the presented algorithms are superior to previous methods at quickly generating near-optimal solutions for small-scale problems, and greatly superior for efficiently finding high-quality solutions for large-scale problems. For example, for problems with a thousand agents and values generated with a uniform distribution, our best approach generates solutions 99.5% of the expected optimal within seconds. For these problems, the state-of-the-art solvers fail to find any feasible solutions at all.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17349

17.Newton Optimization on Helmholtz Decomposition for Continuous Games

Abstract: Many learning problems involve multiple agents optimizing different interactive functions. In these problems, the standard policy gradient algorithms fail due to the non-stationarity of the setting and the different interests of each agent. In fact, algorithms must take into account the complex dynamics of these systems to guarantee rapid convergence towards a (local) Nash equilibrium. In this paper, we propose NOHD (Newton Optimization on Helmholtz Decomposition), a Newton-like algorithm for multi-agent learning problems based on the decomposition of the dynamics of the system in its irrotational (Potential) and solenoidal (Hamiltonian) component. This method ensures quadratic convergence in purely irrotational systems and pure solenoidal systems. Furthermore, we show that NOHD is attracted to stable fixed points in general multi-agent systems and repelled by strict saddle ones. Finally, we empirically compare the NOHD's performance with that of state-of-the-art algorithms on some bimatrix games and continuous Gridworlds environment.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17350

18.Synchronous Dynamical Systems on Directed Acyclic Graphs: Complexity and Algorithms

Abstract: Discrete dynamical systems serve as useful formal models to study diffusion phenomena in social networks. Motivated by applications in systems biology, several recent papers have studied algorithmic and complexity aspects of diffusion problems for dynamical systems whose underlying graphs are directed, and may contain directed cycles. Such problems can be regarded as reachability problems in the phase space of the corresponding dynamical system. We show that computational intractability results for reachability problems hold even for dynamical systems on directed acyclic graphs (dags). We also show that for dynamical systems on dags where each local function is monotone, the reachability problem can be solved efficiently.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17351

19.Evolutionary Game Theory Squared: Evolving Agents in Endogenously Evolving Zero-Sum Games

Abstract: The predominant paradigm in evolutionary game theory and more generally online learning in games is based on a clear distinction between a population of dynamic agents that interact given a fixed, static game. In this paper, we move away from the artificial divide between dynamic agents and static games, to introduce and analyze a large class of competitive settings where both the agents and the games they play evolve strategically over time. We focus on arguably the most archetypal game-theoretic setting---zero-sum games (as well as network generalizations)---and the most studied evolutionary learning dynamic---replicator, the continuous-time analogue of multiplicative weights. Populations of agents compete against each other in a zero-sum competition that itself evolves adversarially to the current population mixture. Remarkably, despite the chaotic coevolution of agents and games, we prove that the

system exhibits a number of regularities. First, the system has conservation laws of an information-theoretic flavor that couple the behavior of all agents and games. Secondly, the system is Poincare recurrent, with effectively all possible initializations of agents and games lying on recurrent orbits that come arbitrarily close to their initial conditions infinitely often. Thirdly, the time-average agent behavior and utility converge to the Nash equilibrium values of the time-average game. Finally, we provide a polynomial time algorithm to efficiently predict this time-average behavior for any such coevolving network game.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17352

20.Value-Decomposition Multi-Agent Actor-Critics

Abstract: The exploitation of extra state information has been an active research area in multi-agent reinforcement learning (MARL). QMIX represents the joint action-value using a non-negative function approximator and achieves the best performance on the StarCraft II micromanagement testbed, a common MARL benchmark. However, our experiments demonstrate that, in some cases, QMIX performs sub-optimally with the A2C framework, a training paradigm that promotes algorithm training efficiency. To obtain a reasonable trade-off between training efficiency and algorithm performance, we extend value-decomposition to actor-critic methods that are compatible with A2C and propose a novel actor-critic framework, value-decomposition actor-critic (VDAC). We evaluate VDAC on the StarCraft II micromanagement task and demonstrate that the proposed framework improves median performance over other actor-critic methods. Furthermore, we use a set of ablation experiments to identify the key factors that contribute to the performance of VDAC.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17353

21.Contract-based Inter-user Usage Coordination in Free-floating Car Sharing

Abstract: We propose a novel distributed user-car matching method based on a contract between users to mitigate the imbalance problem between vehicle distribution and demand in free-floating car sharing. Previous regulation methods involved an incentive system based on the predictions of origin-destination (OD) demand obtained from past usage history. However, the difficulty these methods have in obtaining accurate data limits their applicability. To overcome this drawback, we introduce contract-based coordination among drop-off and pick-up users in which an auction is conducted for drop-off users' intended drop-off locations. We theoretically analyze the proposed method regarding the upper bound of its efficiency. We also compare it with a baseline method and non-regulation scenario on a free-floating car-sharing simulator. The experimental results show that the proposed method achieves a higher social surplus than the existing method.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17354

22.Maintenance of Social Commitments in Multiagent Systems

Abstract: We introduce and formalize a concept of a maintenance commitment, a kind of social commitment characterized by states whose truthhood an agent commits to maintain. This concept of maintenance commitments enables us to capture a richer variety of real-world scenarios than possible using achievement commitments with a temporal condition. By developing a rule-based operational semantics, we study the relationship between agents' achievement and maintenance goals, achievement commitments, and maintenance

commitments. We motivate a notion of coherence which captures alignment between an agents' achievement and maintenance cognitive and social constructs, and prove that, under specified conditions, the goals and commitments of both rational agents individually and of a multiagent system are coherent.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17355

23.Efficient Querying for Cooperative Probabilistic Commitments

Abstract: Multiagent systems can use commitments as the core of a general coordination infrastructure, supporting both cooperative and non-cooperative interactions. Agents whose objectives are aligned, and where one agent can help another achieve greater reward by sacrificing some of its own reward, should choose a cooperative commitment to maximize their joint reward. We present a solution to the problem of how cooperative agents can efficiently find an (approximately) optimal commitment by querying about carefully-selected commitment choices. We prove structural properties of the agents' values as functions of the parameters of the commitment specification, and develop a greedy method for composing a query with provable approximation bounds, which we empirically show can find nearly optimal commitments in a fraction of the time methods that lack our insights require.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17356

24.Coordination Between Individual Agents in Multi-Agent Reinforcement Learning

Abstract: The existing multi-agent reinforcement learning methods (MARL) for determining the coordination between agents focus on either global-level or neighborhood-level coordination between agents. However the problem of coordination between individual agents is remain to be solved. It is crucial for learning an optimal coordinated policy in unknown multi-agent environments to analyze the agent's roles and the correlation between individual agents. To this end, in this paper we propose an agent-level coordination based MARL method. Specifically, it includes two parts in our method. The first is correlation analysis between individual agents based on the Pearson, Spearman, and Kendall correlation coefficients; And the second is an agent-level coordinated training framework where the communication message between weakly correlated agents is dropped out, and a correlation based reward function is built. The proposed method is verified in four mixed cooperative-competitive environments. The experimental results show that the proposed method outperforms the state-of-the-art MARL methods and can measure the correlation between individual agents accurately.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17357

25.Improved Knowledge Modeling and Its Use for Signaling in Multi-Agent Planning with Partial Observability

Abstract: Collaborative Multi-Agent Planning (MAP) problems with uncertainty and partial observability are often modeled as Dec-POMDPs. Yet, in deterministic domains, Qualitative Dec-POMDPs can scale up to much larger problem sizes. The best current QDec solver (QDec-FP) reduces MAP problems to multiple single-agent problems. In this paper, we describe a planner that uses richer information about agents' knowledge to improve upon QDec-FP. With this change, the planner not only scales up to larger problems with more objects, but it can also support signaling, where agents signal information to each other by changing the state of the world.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17420

Reinforce Learning

RL部分收录了：

1. AAAI Technical Track on Application Domains中的7篇文章(1-7)。
2. AAAI Technical Track on Machine Learning I-V中的62篇文章(8-69)。

1.Towered Actor Critic For Handling Multiple Action Types In Reinforcement Learning For Drug Discovery

Abstract: Reinforcement learning (RL) has made significant progress in both abstract and real-world domains, but the majority of state-of-the-art algorithms deal only with monotonic actions. However, some applications require agents to reason over different types of actions. Our application simulates reaction-based molecule generation, used as part of the drug discovery pipeline, and includes both uni-molecular and bi-molecular reactions. This paper introduces a novel framework, towered actor critic (TAC), to handle multiple action types. The TAC framework is general in that it is designed to be combined with any existing RL algorithms for continuous action space. We combine it with TD3 to empirically obtain significantly better results than existing methods in the drug discovery setting. TAC is also applied to RL benchmarks in OpenAI Gym and results show that our framework can improve, or at least does not hurt, performance relative to standard TD3.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16087

2.Queue-Learning: A Reinforcement Learning Approach for Providing Quality of Service

Abstract: End-to-end delay is a critical attribute of quality of service (QoS) in application domains such as cloud computing and computer networks. This metric is particularly important in tandem service systems, where the end-to-end service is provided through a chain of services. Service-rate control is a common mechanism for providing QoS guarantees in service systems. In this paper, we introduce a reinforcement learning-based (RL-based) service-rate controller that provides probabilistic upper-bounds on the end-to-end delay of the system, while preventing the overuse of service resources. In order to have a general framework, we use queueing theory to model the service systems. However, we adopt an RL-based approach to avoid the limitations of queueing-theoretic methods. In particular, we use Deep Deterministic Policy Gradient (DDPG) to learn the service rates (action) as a function of the queue lengths (state) in tandem service systems. In contrast to existing RL-based methods that quantify their performance by the achieved overall reward, which could be hard to interpret or even misleading, our proposed controller provides explicit probabilistic guarantees on the end-to-end delay of the system. The evaluations are presented for a tandem queueing system with non-exponential inter-arrival and service times, the results of which validate our controller's capability in meeting QoS constraints.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16123

3.Content Masked Loss: Human-Like Brush Stroke Planning in a Reinforcement Learning Painting Agent

Abstract: The objective of most Reinforcement Learning painting agents is to minimize the loss between a target image and the paint canvas. Human painter artistry emphasizes important features of the target image rather than simply reproducing it. Using adversarial or L2 losses in the RL painting models, although its final output is generally a work of finesse, produces a stroke sequence that is vastly different from that which a human would produce since the model does not have knowledge about the abstract features in the target image. In order to increase the human-like planning of the model without the use of expensive human data, we introduce a new loss function for use with the model's reward function: Content Masked Loss. In the context of robot painting, Content Masked Loss employs an object detection model to extract features which are used to assign higher weight to regions of the canvas that a human would find important for recognizing content. The results, based on 332 human evaluators, show that the digital paintings produced by our Content Masked model show detectable subject matter earlier in the stroke sequence than existing methods without compromising on the quality of the final painting. Our code is available at https://github.com/pschaldenbrand/ContentMaskedLoss.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16128

4.Commission Fee is not Enough: A Hierarchical Reinforced Framework for Portfolio Management

Abstract: Portfolio management via reinforcement learning is at the forefront of fintech research, which explores how to optimally reallocate a fund into different financial assets over the long term by trial-and-error. Existing methods are impractical since they usually assume each reallocation can be finished immediately and thus ignoring the price slippage as part of the trading cost. To address these issues, we propose a hierarchical reinforced stock trading system for portfolio management (HRPM). Concretely, we decompose the trading process into a hierarchy of portfolio management over trade execution and train the corresponding policies. The high-level policy gives portfolio weights at a lower frequency to maximize the long-term profit and invokes the low-level policy to sell or buy the corresponding shares within a short time window at a higher frequency to minimize the trading cost. We train two levels of policies via a pre-training scheme and an iterative training scheme for data efficiency. Extensive experimental results in the U.S. market and the China market demonstrate that HRPM achieves significant improvement against many state-of-the-art approaches.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16142

5.DeepTrader: A Deep Reinforcement Learning Approach for Risk-Return Balanced Portfolio Management with Market Conditions Embedding

Abstract:Most existing reinforcement learning (RL)-based portfolio management models do not take into account the market conditions, which limits their performance in risk-return balancing. In this paper, we propose DeepTrader, a deep RL method to optimize the investment policy. In particular, to tackle the risk-return balancing problem, our model embeds macro market conditions as an indicator to dynamically adjust the proportion between long and short funds, to lower the risk of market fluctuations, with the negative maximum drawdown as the reward function. Additionally, the model involves a unit to evaluate individual assets, which learns dynamic patterns from historical data with the price rising rate as the reward function. Both temporal and spatial dependencies between assets are captured hierarchically by a specific type of graph structure. Particularly, we find that the estimated causal structure best captures the interrelationships between assets, compared to industry classification and correlation. The two units are complementary and integrated to generate a suitable portfolio which fits the market trend well and strikes a balance between return and risk effectively. Experiments on three well-known stock indexes demonstrate the superiority of DeepTrader in terms of risk-gain criteria.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16144

6.Online 3D Bin Packing with Constrained Deep Reinforcement Learning

Abstract: We solve a challenging yet practically useful variant of 3D Bin Packing Problem (3D-BPP). In our problem, the agent has limited information about the items to be packed into a single bin, and an item must be packed immediately after its arrival without buffering or readjusting. The item's placement also subjects to the constraints of order dependence and physical stability. We formulate this online 3D-BPP as a constrained Markov decision process (CMDP). To solve the problem, we propose an effective and easy-to-implement constrained deep reinforcement learning (DRL) method under the actor-critic framework. In particular, we introduce a prediction-and-projection scheme: The agent first predicts a feasibility mask for the placement actions as an auxiliary task and then uses the mask to modulate the action probabilities output by the actor during training. Such supervision and projection facilitate the agent to learn feasible policies very efficiently. Our method can be easily extended to handle lookahead items, multi-bin packing, and item re-orienting. We have conducted extensive evaluation showing that the learned policy significantly outperforms the state-of-the-art methods. A preliminary user study even suggests that our method might attain a human-level performance.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16155

7.DEAR: Deep Reinforcement Learning for Online Advertising Impression in Recommender Systems

Abstract With the recent prevalence of Reinforcement Learning (RL), there have been tremendous interests in utilizing RL for online advertising in recommendation platforms (e.g., e-commerce and news feed sites). However, most RL-based advertising algorithms focus on optimizing ads' revenue while ignoring the possible negative influence of ads on user experience of recommended items (products, articles and videos). Developing an optimal advertising algorithm in recommendations faces immense challenges because interpolating ads improperly or too frequently may decrease user experience, while interpolating fewer ads will reduce the advertising revenue. Thus, in this paper, we propose a novel advertising strategy for the rec/ads trade-off. To be specific, we develop an RL-based framework that can continuously update its advertising strategies and maximize reward in the long run. Given a recommendation list, we design a novel Deep Q-network architecture that can determine three internally related tasks jointly, i.e., (i) whether to interpolate an ad or not in the recommendation list, and if yes, (ii) the optimal ad and (iii) the optimal location to interpolate. The experimental results based on real-world data demonstrate the effectiveness of the proposed framework.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16156

8.Improved Worst-Case Regret Bounds for Randomized Least-Squares Value Iteration

Abstract This paper studies regret minimization with randomized value functions in reinforcement learning. In tabular finite-horizon Markov Decision Processes, we introduce a clipping variant of one classical Thompson Sampling (TS)-like algorithm, randomized least-squares value iteration (RLSVI). Our $\tilde{\mathrm{O}}(H^2S\sqrt{AT})$ high-probability worst-case regret bound improves the previous sharpest worst-case regret bounds for RLSVI and matches the existing state-of-the-art worst-case TS-based regret bounds.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16813

9.Deep Bayesian Quadrature Policy Optimization

Abstract: We study the problem of obtaining accurate policy gradient estimates using a finite number of samples. Monte-Carlo methods have been the default choice for policy gradient estimation, despite suffering from high variance in the gradient estimates. On the other hand, more sample efficient alternatives like Bayesian quadrature methods have received little attention due to their high computational complexity. In this work, we propose deep Bayesian quadrature policy gradient (DBQPG), a computationally efficient high-dimensional generalization of Bayesian quadrature, for policy gradient estimation. We show that DBQPG can substitute Monte-Carlo estimation in policy gradient methods, and demonstrate its effectiveness on a set of continuous control benchmarks. In comparison to Monte-Carlo estimation, DBQPG provides (i) more accurate gradient estimates with a significantly lower variance, (ii) a consistent improvement in the sample complexity and average return for several deep policy gradient algorithms, and, (iii) the uncertainty in gradient estimation that can be incorporated to further improve the performance.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16817

10.Learning Task-Distribution Reward Shaping with Meta-Learning

Abstract: Reward shaping is one of the most effective methods to tackle the crucial yet challenging problem of credit assignment and accelerate Reinforcement Learning. However, designing shaping functions usually requires rich expert knowledge and hand-engineering, and the difficulties are further exacerbated given multiple tasks to solve. In this paper, we consider reward shaping on a distribution of tasks that share state spaces but not necessarily action spaces. We provide insights into optimal reward shaping, and propose a novel meta-learning framework to automatically learn such reward shaping to apply on newly sampled tasks. Theoretical analysis and extensive experiments establish us as the state-of-the-art in learning task-distribution reward shaping, outperforming previous such works (Konidaris and Barto 2006; Snel and Whiteson 2014). We further show that our method outperforms learning intrinsic rewards (Yang et al. 2019; Zheng et al. 2020), outperforms Rainbow (Hessel et al. 2018) in complex pixel-based CoinRun games, and is also better than hand-designed reward shaping on grids. While the goal of this paper is to learn reward shaping rather than to propose new general meta-learning algorithms as PEARL (Rakelly et al. 2019) or MQL (Fakoor et al. 2020), our framework based on MAML (Finn, Abbeel, and Levine 2017) also outperforms PEARL / MQL, and could combine with them for further improvement.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17337

11.An Enhanced Advising Model in Teacher-Student Framework using State Categorization

Abstract: The teacher-student framework aims to improve the sample efficiency of RL algorithms by deploying an advising mechanism in which a teacher helps a student by guiding its exploration. Prior work in this field has considered an advising mechanism where the teacher advises the student about the optimal action to take in a given state. However, real-world teachers can leverage domain expertise to provide more informative signals. Using this insight, we propose to extend the current advising framework wherein the teacher would provide not only the optimal action but also a qualitative assessment of the state. We introduce a novel architecture, namely Advice Replay Memory (ARM), to effectively reuse the advice provided by the teacher. We demonstrate the robustness of our approach by showcasing our experiments on multiple Atari 2600 games using a fixed set of hyper-parameters. Additionally, we show that a student taking help even from a sub-optimal teacher can achieve significant performance boosts and eventually outperform the teacher. Our approach outperforms the baselines even when provided with comparatively suboptimal teachers and an advising budget, which is smaller by orders of magnitude. The contributions of our paper are 4-fold (a) effectively leveraging a teacher's knowledge by richer advising (b) introduction of ARM to effectively reuse the advice throughout learning (c) ability to achieve significant performance boost even with a coarse state categorization (d) enabling the student to outperform the teacher.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16823

13.Deep Radial-Basis Value Functions for Continuous Control

Abstract: A core operation in reinforcement learning (RL) is finding an action that is optimal with respect to a learned value function. This operation is often challenging when the learned value function takes continuous actions as input. We introduce deep radial-basis value functions (RBVFs): value functions learned using a deep network with a radial-basis function (RBF) output layer. We show that the maximum action-value with respect to a deep RBVF can be approximated easily and accurately. Moreover, deep RBVFs can represent any true value function owing to their support for universal function approximation. We extend the standard DQN algorithm to continuous control by endowing the agent with a deep RBVF. We show that the resultant agent, called RBF-DQN, significantly outperforms value-function-only baselines, and is competitive with state-of-the-art actor-critic algorithms.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16828

14.Relative Variational Intrinsic Control

Abstract: In the absence of external rewards, agents can still learn useful behaviors by identifying and mastering a set of diverse skills within their environment. Existing skill learning methods use mutual information objectives to incentivize each skill to be diverse and distinguishable from the rest. However, if care is not taken to constrain the ways in which the skills are diverse, trivially diverse skill sets can arise. To ensure useful skill diversity, we propose a novel skill learning objective, Relative Variational Intrinsic Control (RVIC), which incentivizes learning skills that are distinguishable in how they change the agent's relationship to its environment. The resulting set of skills tiles the space of affordances available to the agent. We qualitatively analyze skill behaviors on multiple environments and show how RVIC skills are more useful than skills discovered by existing methods in hierarchical reinforcement learning.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16832

15.Sample-Specific Output Constraints for Neural Networks

Abstract: It is common practice to constrain the output space of a neural network with the final layer to a problem-specific value range. However, for many tasks it is desired to restrict the output space for each input independently to a different subdomain with a non-trivial geometry, e.g. in safety-critical applications, to exclude hazardous outputs sample-wise. We propose ConstraintNet—a scalable neural network architecture which constrains the output space in each forward pass independently. Contrary to prior approaches, which perform a projection in the final layer, ConstraintNet applies an input-dependent parametrization of the constrained output space. Thereby, the complete interior of the constrained region is covered and computational costs are reduced significantly. For constraints in form of convex polytopes, we leverage the vertex representation to specify the parametrization. The second modification consists of adding an auxiliary input in form of a tensor description of the constraint to enable the handling of multiple constraints for the same sample. Finally, ConstraintNet is end-to-end trainable with almost no overhead in the forward and backward pass. We demonstrate ConstraintNet on two regression tasks: First, we modify a CNN and construct several constraints for facial landmark detection tasks. Second, we demonstrate the application to a follow object controller for vehicles and accomplish safe reinforcement learning in this case. In both experiments, ConstraintNet improves performance and we conclude that our approach is promising for applying neural networks in safety-critical environments.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16841

16.High-Confidence Off-Policy (or Counterfactual) Variance Estimation

Abstract: Many sequential decision-making systems leverage data collected using prior policies to propose a new policy. For critical applications, it is important that high-confidence guarantees on the new policy's behavior are provided before deployment, to ensure that the policy will behave as desired. Prior works have studied high-confidence off-policy estimation of the expected return, however, high-confidence off-policy estimation of the variance of returns can be equally critical for high-risk applications. In this paper we tackle the previously open problem of estimating and bounding, with high confidence, the variance of returns from off-policy data.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16855

17.Addressing Action Oscillations through Learning Policy Inertia

Abstract: Deep reinforcement learning (DRL) algorithms have been demonstrated to be effective on a wide range of challenging decision making and control tasks. However, these methods typically suffer from severe action oscillations in particular in discrete action setting, which means that agents select different actions within consecutive steps even though states only slightly differ. This issue is often neglected since we usually evaluate the quality of a policy using cumulative rewards only. Action oscillation strongly affects the user experience and even causes serious potential security menace especially in real-world domains with the main concern of safety, such as autonomous driving. In this paper, we introduce Policy Inertia Controller (PIC) which serves as a generic plug-in framework to off-the-shelf DRL algorithms, to enable adaptive balance between the optimality and smoothness in a formal way. We propose Nested Policy Iteration as a general training algorithm for PIC-augmented policy which ensures monotonically non-decreasing updates.Further, we derive a practical DRL algorithm, namely Nested Soft Actor-Critic. Experiments on a collection of autonomous driving tasks and several Atari games suggest that our approach demonstrates substantial oscillation reduction than a range of commonly adopted baselines with almost no performance degradation.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16864

18.Transfer Learning for Efficient Iterative Safety Validation

Abstract: Safety validation is important during the development of safety-critical autonomous systems but can require significant computational effort. Existing algorithms often start from scratch each time the system under test changes. We apply transfer learning to improve the efficiency of reinforcement learning based safety validation algorithms when applied to related systems. Knowledge from previous safety validation tasks is encoded through the action value function and transferred to future tasks with a learned set of attention weights. Including a learned state and action value transformation for each source task can improve performance even when systems have substantially different failure modes. We conduct experiments on safety validation tasks in gridworld and autonomous driving scenarios. We show that transfer learning can improve the initial and final performance of validation algorithms and reduce the number of training steps.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16876

19.The Value-Improvement Path: Towards Better Representations for Reinforcement Learning

Abstract: In value-based reinforcement learning (RL), unlike in supervised learning, the agent faces not a single, stationary, approximation problem, but a sequence of value prediction problems. Each time the policy improves, the nature of the problem changes, shifting both the distribution of states and their values. In this paper we take a novel perspective, arguing that the value prediction problems faced by an RL agent should not be addressed in isolation, but rather as a single, holistic, prediction problem. An RL algorithm generates a sequence of policies that, at least approximately, improve towards the optimal policy. We explicitly characterize the associated sequence of value functions and call it the value-improvement path. Our main idea is to approximate the value-improvement path holistically, rather than to solely track the value function of the current policy. Specifically, we discuss the impact that this holistic view of RL has on representation learning. We demonstrate that a representation that spans the past value-improvement path will also provide an accurate value approximation for future policy improvements. We use this insight to better understand existing approaches to auxiliary tasks and to propose new ones. To test our hypothesis empirically, we augmented a standard deep RL agent with an auxiliary task of learning the value-improvement path. In a study of Atari 2600 games, the augmented agent achieved approximately double the mean and median performance of the baseline agent.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/16880

20.Loop Estimator for Discounted Values in Markov Reward Processes

Abstract: At the working heart of policy iteration algorithms commonly used and studied in the discounted setting of reinforcement learning, the policy evaluation step estimates the value of states with samples from a Markov reward process induced by following a Markov policy in a Markov decision process. We propose a simple and efficient estimator called loop estimator that exploits the regenerative structure of Markov reward processes without explicitly estimating a full model. Our method enjoys a space complexity of $O(1)$when estimating the value of a single positive recurrent state s unlike TD with $O(S)$ or model-based methods with $O(S^2)$. Moreover, the regenerative structure enables us to show, without relying on the generative model approach, that the estimator has an instance-dependent convergence rate of $O\left(\sqrt{\tau_s/T}\right)$over steps T on a single sample path, where τ_s is the maximal expected hitting time to state s. In preliminary numerical experiments, the loop estimator outperforms model-free methods, such as $TD(k)$, and is competitive with the model-based estimator.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16881

21.Reinforcement Learning with Trajectory Feedback

Abstract: The standard feedback model of reinforcement learning requires revealing the reward of every visited state-action pair. However, in practice, it is often the case that such frequent feedback is not available. In this work, we take a first step towards relaxing this assumption and require a weaker form of feedback, which we refer to as $trajectory\ feedback$. Instead of observing the reward obtained after every action, we assume we only receive a score that represents the quality of the whole trajectory observed by the agent, namely, the sum of all rewards obtained over this trajectory. We extend reinforcement learning algorithms to this setting, based on least-squares estimation of the unknown reward, for both the known and unknown transition model cases, and study the performance of these algorithms by analyzing their regret. For cases where the transition model is unknown, we offer a hybrid optimistic-Thompson Sampling approach that results in a tractable algorithm.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16895

22.Towards Effective Context for Meta-Reinforcement Learning: an Approach based on Contrastive Learning

Abstract: Context, the embedding of previous collected trajectories, is a powerful construct for Meta-Reinforcement Learning (Meta-RL) algorithms. By conditioning on an effective context, Meta-RL policies can easily generalize to new tasks within a few adaptation steps. We argue that improving the quality of context involves answering two questions: 1. How to train a compact and sufficient encoder that can embed the task-specific information contained in prior trajectories? 2. How to collect informative trajectories of which the corresponding context reflects the specification of tasks? To this end, we propose a novel Meta-RL framework called CCM (Contrastive learning augmented Context-based Meta-RL). We first focus on the contrastive nature behind different tasks and leverage it to train a compact and sufficient context encoder. Further, we train a separate exploration policy and theoretically derive a new information-gain-based objective which aims to collect informative trajectories in a few steps. Empirically, we evaluate our approaches on common benchmarks as well as several complex sparse-reward environments. The experimental results show that CCM outperforms state-of-the-art algorithms by addressing previously mentioned problems respectively.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16914

23.Stabilizing Q Learning Via Soft Mellowmax Operator

Abstract: Learning complicated value functions in high dimensional state space by function approximation is a challenging task, partially due to that the max-operator used in temporal difference updates can theoretically cause instability for most linear or non-linear approximation schemes. Mellowmax is a recently proposed differentiable and non-expansion softmax operator that allows a convergent behavior in learning and planning. Unfortunately, the performance bound for the fixed point it converges to remains unclear, and in practice, its parameter is sensitive to various domains and has to be tuned case by case. Finally, the Mellowmax operator may suffer from oversmoothing as it ignores the probability being taken for each action when aggregating them. In this paper we address all the above issues with an enhanced Mellowmax operator, named SM2 (Soft Mellowmax). Particularly, the proposed operator is reliable, easy to implement, and has provable performance guarantee, while preserving all the advantages of Mellowmax. Furthermore, we show that our SM2 operator can be applied to the challenging multi-agent reinforcement learning scenarios, leading to stable value function approximation and state of the art performance.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16919

24.DeepSynth: Automata Synthesis for Automatic Task Segmentation in Deep Reinforcement Learning

Abstract: This paper proposes DeepSynth, a method for effective training of deep Reinforcement Learning (RL) agents when the reward is sparse and non-Markovian, but at the same time progress towards the reward requires achieving an unknown sequence of high-level objectives. Our method employs a novel algorithm for synthesis of compact automata to uncover this sequential structure automatically. We synthesise a human-interpretable automaton from trace data collected by exploring the environment. The state space of the environment is then enriched with the synthesised automaton so that the generation of a control policy by deep RL is guided by the discovered structure encoded in the automaton. The proposed approach is able to cope with both high-dimensional, low-level features and unknown sparse non-Markovian rewards. We have evaluated DeepSynth's performance in a set of experiments that includes the Atari game Montezuma's Revenge. Compared to existing approaches, we obtain a reduction of two orders of magnitude in the number of iterations required for policy synthesis, and also a significant improvement in scalability.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16935

25.Learning with Safety Constraints: Sample Complexity of Reinforcement Learning for Constrained MDPs

Abstract: Many physical systems have underlying safety considerations that require that the policy employed ensures the satisfaction of a set of constraints. The analytical formulation usually takes the form of a Constrained Markov Decision Process (CMDP). We focus on the case where the CMDP is unknown, and RL algorithms obtain samples to discover the model and compute an optimal constrained policy. Our goal is to characterize the relationship between safety constraints and the number of samples needed to ensure a desired level of accuracy---both objective maximization and constraint satisfaction---in a PAC sense. We explore two classes of RL algorithms, namely, (i) a generative model based approach, wherein samples are taken initially to estimate a model, and (ii) an online approach, wherein the model is updated as samples are obtained. Our main finding is that compared to the best known bounds of the unconstrained regime, the sample complexity of constrained RL algorithms are increased by a factor that is logarithmic in the number of constraints, which suggests that the approach may be easily utilized in real systems.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16937

26.Learning to Reweight Imaginary Transitions for Model-Based Reinforcement Learning

Abstract: Model-based reinforcement learning (RL) is more sample efficient than model-free RL by using imaginary trajectories generated by the learned dynamics model. When the model is inaccurate or biased, imaginary trajectories may be deleterious for training the action-value and policy functions. To alleviate such problem, this paper proposes to adaptively reweight the imaginary transitions, so as to reduce the negative effects of poorly generated trajectories. More specifically, we evaluate the effect of an imaginary transition by calculating the change of the loss computed on the real samples when we use the transition to train the action-value and policy functions. Based on this evaluation criterion, we construct the idea of reweighting each imaginary transition by a well-designed meta-gradient algorithm. Extensive experimental results demonstrate that our method outperforms state-of-the-art model-based and model-free RL algorithms on multiple tasks. Visualization of our changing weights further validates the necessity of utilizing reweight scheme.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16958

27.Variance Penalized On-Policy and Off-Policy Actor-Critic

Abstract: Reinforcement learning algorithms are typically geared towards optimizing the expected return of an agent. However, in many practical applications, low variance in the return is desired to ensure the reliability of an algorithm. In this paper, we propose on-policy and off-policy actor-critic algorithms that optimize a performance criterion involving both mean and variance in the return. Previous work uses the second moment of return to estimate the variance indirectly. Instead, we use a much simpler recently proposed direct variance estimator which updates the estimates incrementally using temporal difference methods. Using the variance-penalized criterion, we guarantee the convergence of our algorithm to locally optimal policies for finite state action Markov decision processes. We demonstrate the utility of our algorithm in tabular and continuous MuJoCo domains. Our approach not only performs on par with actor-critic and prior variance-penalization baselines in terms of expected return, but also generates trajectories which have lower variance in the return.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16964

28.Action Candidate Based Clipped Double Q-learning for Discrete and Continuous Action Tasks

Abstract: Double Q-learning is a popular reinforcement learning algorithm in Markov decision process (MDP) problems. Clipped Double Q-learning, as an effective variant of Double Q-learning, employs the clipped double estimator to approximate the maximum expected action value. Due to the underestimation bias of the clipped double estimator, performance of clipped Double Q-learning may be degraded in some stochastic environments. In this paper, in order to reduce the underestimation bias, we propose an action candidate based clipped double estimator for Double Q-learning. Specifically, we first select a set of elite action candidates with the high action values from one set of estimators. Then, among these candidates, we choose the highest valued action from the other set of estimators. Finally, we use the maximum value in the second set of estimators to clip the action value of the chosen action in the first set of estimators and the clipped value is used for approximating the maximum expected action value. Theoretically, the underestimation bias in our clipped Double Q-learning decays monotonically as the number of the action candidates decreases. Moreover, the number of action candidates controls the trade-off between the overestimation and underestimation biases. In addition, we also extend our

clipped Double Q-learning to continuous action tasks via approximating the elite continuous action candidates. We empirically verify that our algorithm can more accurately estimate the maximum expected action value on some toy environments and yield good performance on several benchmark problems.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16973

29.Temporal-Logic-Based Reward Shaping for Continuing Reinforcement Learning Tasks

Abstract: In continuing tasks, average-reward reinforcement learning may be a more appropriate problem formulation than the more common discounted reward formulation. As usual, learning an optimal policy in this setting typically requires a large amount of training experiences. Reward shaping is a common approach for incorporating domain knowledge into reinforcement learning in order to speed up convergence to an optimal policy. However, to the best of our knowledge, the theoretical properties of reward shaping have thus far only been established in the discounted setting. This paper presents the first reward shaping framework for average-reward learning and proves that, under standard assumptions, the optimal policy under the original reward function can be recovered. In order to avoid the need for manual construction of the shaping function, we introduce a method for utilizing domain knowledge expressed as a temporal logic formula. The formula is automatically translated to a shaping function that provides additional reward throughout the learning process. We evaluate the proposed method on three continuing tasks. In all cases, shaping speeds up the average-reward learning rate without any reduction in the performance of the learned policy compared to relevant baselines.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16975

30.A Sample-Efficient Algorithm for Episodic Finite-Horizon MDP with Constraints

Abstract: Constrained Markov decision processes (CMDPs) formalize sequential decision-making problems whose objective is to minimize a cost function while satisfying constraints on various cost functions. In this paper, we consider the setting of episodic fixed-horizon CMDPs. We propose an online algorithm which leverages the linear programming formulation of repeated optimistic planning for finite-horizon CMDP to provide a probably approximately correctness (PAC) guarantee on the number of episodes needed to ensure a near optimal policy, i.e., with resulting objective value close to that of the optimal value and satisfying the constraints within low tolerance, with high probability. The number of episodes needed is shown to have linear dependence on the sizes of the state and action spaces and quadratic dependence on the time horizon and an upper bound on the number of possible successor states for a state-action pair. Therefore, if the upper bound on the number of possible successor states is much smaller than the size of the state space, the number of needed episodes becomes linear in the sizes of the state and action spaces and quadratic in the time horizon.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16979

31.Exploration via State influence Modeling

Abstract: This paper studies the challenging problem of reinforcement learning (RL) in hard exploration tasks with sparse rewards. It focuses on the exploration stage before the agent gets the first positive reward, in which case, traditional RL algorithms with simple exploration strategies often work poorly. Unlike previous methods using some attribute of a single state as the intrinsic reward to encourage exploration, this work leverages the social influence between different states to permit more efficient exploration. It introduces a general intrinsic reward construction method to evaluate the social influence of states dynamically. Three kinds of social influence are introduced for a state: conformity, power, and authority. By measuring the state's social influence, agents quickly find the focus state during the exploration process. The proposed RL framework with state social influence evaluation works well in hard exploration task. Extensive experimental analyses and comparisons in Grid Maze and many hard exploration Atari 2600 games demonstrate its high exploration efficiency.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/16981

32.Metrics and Continuity in Reinforcement Learning

Abstract: In most practical applications of reinforcement learning, it is untenable to maintain direct estimates for individual states; in continuous-state systems, it is impossible. Instead, researchers often leverage {\em state similarity} (whether explicitly or implicitly) to build models that can generalize well from a limited set of samples. The notion of state similarity used, and the neighbourhoods and topologies they induce, is thus of crucial importance, as it will directly affect the performance of the algorithms. Indeed, a number of recent works introduce algorithms assuming the existence of "well-behaved" neighbourhoods, but leave the full specification of such topologies for future work. In this paper we introduce a unified formalism for defining these topologies through the lens of metrics. We establish a hierarchy amongst these metrics and demonstrate their theoretical implications on the Markov Decision Process specifying the reinforcement learning problem. We complement our theoretical results with empirical evaluations showcasing the differences between the metrics considered.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17005

33.Lipschitz Lifelong Reinforcement Learning

Abstract: We consider the problem of knowledge transfer when an agent is facing a series of Reinforcement Learning (RL) tasks. We introduce a novel metric between Markov Decision Processes and establish that close MDPs have close optimal value functions. Formally, the optimal value functions are Lipschitz continuous with respect to the tasks space. These theoretical results lead us to a value-transfer method for Lifelong RL, which we use to build a PAC-MDP algorithm with improved convergence rate. Further, we show the method to experience no negative transfer with high probability. We illustrate the benefits of the method in Lifelong RL experiments.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17006

34.Bayesian Distributional Policy Gradients

Abstract: Distributional Reinforcement Learning (RL) maintains the entire probability distribution of the reward-to-go, i.e. the return, providing more learning signals that account for the uncertainty associated with policy performance, which may be beneficial for trading off exploration and exploitation and policy learning in general. Previous works in distributional RL

focused mainly on computing the state-action-return distributions, here we model the state-return distributions. This enables us to translate successful conventional RL algorithms that are based on state values into distributional RL. We formulate the distributional Bellman operation as an inference-based auto-encoding process that minimises Wasserstein metrics between target/model return distributions. The proposed algorithm, BDPG (Bayesian Distributional Policy Gradients), uses adversarial training in joint-contrastive learning to estimate a variational posterior from the returns. Moreover, we can now interpret the return prediction uncertainty as an information gain, which allows to obtain a new curiosity measure that helps BDPG steer exploration actively and efficiently. We demonstrate in a suite of Atari 2600 games and MuJoCo tasks, including well known hard-exploration challenges, how BDPG learns generally faster and with higher asymptotic performance than reference distributional RL algorithms.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17024

35.Online Optimal Control with Affine Constraints

Abstract: This paper considers online optimal control with affine constraints on the states and actions under linear dynamics with bounded random disturbances. The system dynamics and constraints are assumed to be known and time invariant but the convex stage cost functions change adversarially. To solve this problem, we propose Online Gradient Descent with Buffer Zones (OGD-BZ). Theoretically, we show that OGD-BZ with proper parameters can guarantee the system to satisfy all the constraints despite any admissible disturbances. Further, we investigate the policy regret of OGD-BZ, which compares OGD-BZ's performance with the performance of the optimal linear policy in hindsight. We show that OGD-BZ can achieve a policy regret upper bound that is square root of the horizon length multiplied by some logarithmic terms of the horizon length under proper algorithm parameters.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17035

36.An Efficient Algorithm for Deep Stochastic Contextual Bandits

Abstract: In stochastic contextual bandit (SCB) problems, an agent selects an action based on certain observed context to maximize the cumulative reward over iterations. Recently there have been a few studies using a deep neural network (DNN) to predict the expected reward for an action, and the DNN is trained by a stochastic gradient based method. However, convergence analysis has been greatly ignored to examine whether and where these methods converge. In this work, we formulate the SCB that uses a DNN reward function as a non-convex stochastic optimization problem, and design a stage-wised stochastic gradient descent algorithm to optimize the problem and determine the action policy. We prove that with high probability, the action sequence chosen by our algorithm converges to a greedy action policy respecting a local optimal reward function. Extensive experiments have been performed to demonstrate the effectiveness and efficiency of the proposed algorithm on multiple real-world datasets.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17335

37.Exact Reduction of Huge Action Spaces in General Reinforcement Learning

Abstract: The reinforcement learning (RL) framework formalizes the notion of learning with interactions. Many real-world problems have large state-spaces and/or action-spaces such as in Go, StarCraft, protein folding, and robotics or are non-Markovian, which cause significant challenges to RL algorithms. In this work we address the large action-space problem by sequentializing actions, which can reduce the action-space size significantly, even down to two actions at the expense of an increased planning horizon. We provide explicit and exact constructions and equivalence proofs for all quantities of interest for arbitrary history-based processes. In the case of MDPs, this could help RL algorithms that bootstrap. In this work we show how action-binarization in the non-MDP case can significantly improve Extreme State Aggregation (ESA) bounds. ESA allows casting any (non-MDP, non-ergodic, history-based) RL problem into a fixed-sized non-Markovian state-space with the help of a surrogate Markovian process. On the upside, ESA enjoys similar optimality guarantees as Markovian models do. But a downside is that the size of the aggregated state-space becomes exponential in the size of the action-space. In this work, we patch this issue by binarizing the action-space. We provide an upper bound on the number of states of this binarized ESA that is logarithmic in the original action-space size, a double-exponential improvement.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17074

38.Policy Optimization as Online Learning with Mediator Feedback

Abstract: Policy Optimization (PO) is a widely used approach to address continuous control tasks. In this paper, we introduce the notion of mediator feedback that frames PO as an online learning problem over the policy space. The additional available information, compared to the standard bandit feedback, allows reusing samples generated by one policy to estimate the performance of other policies. Based on this observation, we propose an algorithm, RANDomized-exploration policy Optimization via Multiple Importance Sampling with Truncation (RANDOMIST), for regret minimization in PO, that employs a randomized exploration strategy, differently from the existing optimistic approaches. When the policy space is finite, we show that under certain circumstances, it is possible to achieve constant regret, while always enjoying logarithmic regret. We also derive problem-dependent regret lower bounds. Then, we extend RANDOMIST to compact policy spaces. Finally, we provide numerical simulations on finite and compact policy spaces, in comparison with PO and bandit baselines.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17083

39.Scheduling of Time-Varying Workloads Using Reinforcement Learning

Abstract: Resource usage of production workloads running on shared compute clusters often fluctuate significantly across time. While simultaneous spike in the resource usage between two workloads running on the same machine can create performance degradation, unused resources in a machine results in wastage and undesirable operational characteristics for a compute cluster. Prior works did not consider such temporal resource fluctuations or their alignment for scheduling decisions. Due to the variety of time-varying workloads, their complex resource usage characteristics, it is challenging to design well-defined heuristics for scheduling them optimally across different machines in a cluster. In this paper, we propose a Deep Reinforcement Learning (DRL) based approach to exploit various temporal resource usage patterns of time varying workloads as well as a technique for creating equivalence classes among a large number of production workloads to improve scalability of our method. Validations with real production

traces from Google and Alibaba show that our technique can significantly improve metrics for operational excellence (e.g. utilization, fragmentation, resource exhaustion etc.) for a cluster, compared to the baselines.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17088

40.Advice-Guided Reinforcement Learning in a non-Markovian Environment

Abstract: We study a class of reinforcement learning tasks in which the agent receives its reward for complex, temporally-extended behaviors sparsely. For such tasks, the problem is how to augment the state-space so as to make the reward function Markovian in an efficient way. While some existing solutions assume that the reward function is explicitly provided to the learning algorithm (e.g., in the form of a reward machine), the others learn the reward function from the interactions with the environment, assuming no prior knowledge provided by the user. In this paper, we generalize both approaches and enable the user to give advice to the agent, representing the user's best knowledge about the reward function, potentially fragmented, partial, or even incorrect. We formalize advice as a set of DFAs and present a reinforcement learning algorithm that takes advantage of such advice, with optimal con- vergence guarantee. The experiments show that using well- chosen advice can reduce the number of training steps needed for convergence to optimal policy, and can decrease the computation time to learn the reward function by up to two orders of magnitude.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17096

41.Distributional Reinforcement Learning via Moment Matching

Abstract: We consider the problem of learning a set of probability distributions from the empirical Bellman dynamics in distributional reinforcement learning (RL), a class of state-of-the-art methods that estimate the distribution, as opposed to only the expectation, of the total return. We formulate a method that learns a finite set of statistics from each return distribution via neural networks, as in the distributional RL literature. Existing distributional RL methods however constrain the learned statistics to predefined functional forms of the return distribution which is both restrictive in representation and difficult in maintaining the predefined statistics. Instead, we learn unrestricted statistics, i.e., deterministic (pseudo-)samples, of the return distribution by leveraging a technique from hypothesis testing known as maximum mean discrepancy (MMD), which leads to a simpler objective amenable to backpropagation. Our method can be interpreted as implicitly matching all orders of moments between a return distribution and its Bellman target. We establish sufficient conditions for the contraction of the distributional Bellman operator and provide finite-sample analysis for the deterministic samples in distribution approximation. Experiments on the suite of Atari games show that our method outperforms the standard distributional RL baselines and sets a new record in the Atari games for non-distributed agents.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17104

42.Inverse Reinforcement Learning From Like-Minded Teachers

Abstract: We study the problem of learning a policy in a Markov decision process (MDP) based on observations of the actions taken by multiple teachers. We assume that the teachers are like-minded in that their reward functions -- while different from each other -- are random perturbations of an underlying reward function. Under this assumption, we demonstrate that inverse reinforcement learning algorithms that satisfy a certain property -- that of matching

feature expectations -- yield policies that are approximately optimal with respect to the underlying reward function, and that no algorithm can do better in the worst case. We also show how to efficiently recover the optimal policy when the MDP has one state -- a setting that is akin to multi-armed bandits.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17110

43.Multinomial Logit Contextual Bandits: Provable Optimality and Practicality

Abstract: We consider a sequential assortment selection problem where the user choice is given by a multinomial logit (MNL) choice model whose parameters are unknown. In each period, the learning agent observes a d-dimensional contextual information about the user and the N available items, and offers an assortment of size K to the user, and observes the bandit feedback of the item chosen from the assortment. We propose upper confidence bound based algorithms for this MNL contextual bandit. The first algorithm is a simple and practical method that achieves an O(d√T) regret over T rounds. Next, we propose a second algorithm which achieves a O(√dT) regret. This matches the lower bound for the MNL bandit problem, up to logarithmic terms, and improves on the best-known result by a √d factor. To establish this sharper regret bound, we present a non-asymptotic confidence bound for the maximum likelihood estimator of the MNL model that may be of independent interest as its own theoretical contribution. We then revisit the simpler, significantly more practical, first algorithm and show that a simple variant of the algorithm achieves the optimal regret for a broad class of important applications.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17111

44.Robust Reinforcement Learning: A Case Study in Linear Quadratic Regulation

Abstract: This paper studies the robustness of reinforcement learning algorithms to errors in the learning process. Specifically, we revisit the benchmark problem of discrete-time linear quadratic regulation (LQR) and study the long-standing open question: Under what conditions is the policy iteration method robustly stable from a dynamical systems perspective? Using advanced stability results in control theory, it is shown that policy iteration for LQR is inherently robust to small errors in the learning process and enjoys small-disturbance input-to-state stability: whenever the error in each iteration is bounded and small, the solutions of the policy iteration algorithm are also bounded, and, moreover, enter and stay in a small neighborhood of the optimal LQR solution. As an application, a novel off-policy optimistic least-squares policy iteration for the LQR problem is proposed, when the system dynamics are subjected to additive stochastic disturbances. The proposed new results in robust reinforcement learning are validated by a numerical example.

HomePage： https://ojs.aaai.org/index.php/AAAI/article/view/17122

45.Uncertainty-Aware Policy Optimization: A Robust, Adaptive Trust Region Approach

Abstract: In order for reinforcement learning techniques to be useful in real-world decision making processes, they must be able to produce robust performance from limited data. Deep policy optimization methods have achieved impressive results on complex tasks, but their real-world adoption remains limited because they often require significant amounts of data to succeed. When combined with small sample sizes, these methods can result in unstable learning

due to their reliance on high-dimensional sample-based estimates. In this work, we develop techniques to control the uncertainty introduced by these estimates. We leverage these techniques to propose a deep policy optimization approach designed to produce stable performance even when data is scarce. The resulting algorithm, Uncertainty-Aware Trust Region Policy Optimization, generates robust policy updates that adapt to the level of uncertainty present throughout the learning process.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17130

46.Visual Transfer For Reinforcement Learning Via Wasserstein Domain Confusion

Abstract: We introduce Wasserstein Adversarial Proximal Policy Optimization (WAPPO), a novel algorithm for visual transfer in Reinforcement Learning that explicitly learns to align the distributions of extracted features between a source and target task. WAPPO approximates and minimizes the Wasserstein-1 distance between the distributions of features from source and target domains via a novel Wasserstein Confusion objective. WAPPO outperforms the prior state-of-the-art in visual transfer and successfully transfers policies across Visual Cartpole and both the easy and hard settings of of 16 OpenAI Procgen environments.

https://ojs.aaai.org/index.php/AAAI/article/view/17139

47.Inverse Reinforcement Learning with Explicit Policy Estimates

Abstract: Various methods for solving the inverse reinforcement learning (IRL) problem have been developed independently in machine learning and economics. In particular, the method of Maximum Causal Entropy IRL is based on the perspective of entropy maximization, while related advances in the field of economics instead assume the existence of unobserved action shocks to explain expert behavior (Nested Fixed Point Algorithm, Conditional Choice Probability method, Nested Pseudo-Likelihood Algorithm). In this work, we make previously unknown connections between these related methods from both fields. We achieve this by showing that they all belong to a class of optimization problems, characterized by a common form of the objective, the associated policy and the objective gradient. We demonstrate key computational and algorithmic differences which arise between the methods due to an approximation of the optimal soft value function, and describe how this leads to more efficient algorithms. Using insights which emerge from our study of this class of optimization problems, we identify various problem scenarios and investigate each method's suitability for these problems.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17141

48.Theoretically Principled Deep RL Acceleration via Nearest Neighbor Function Approximation

Abstract: Recently, deep reinforcement learning (RL) has achieved remarkable empirical success by integrating deep neural networks into RL frameworks. However, these algorithms often require a large number of training samples and admit little theoretical understanding. To mitigate these issues, we propose a theoretically principled nearest neighbor (NN) function approximator that can replace the value networks in deep RL methods. Inspired by human similarity judgments, the NN approximator estimates the action values using rollouts on past observations and can provably obtain a small regret bound that depends only on the intrinsic complexity of the environment. We present (1) Nearest Neighbor Actor-Critic (NNAC), an online policy gradient algorithm that demonstrates the practicality of combining function approximation with deep RL,

and (2) a plug-and-play NN update module that aids the training of existing deep RL methods. Experiments on classical control and MuJoCo locomotion tasks show that the NN-accelerated agents achieve higher sample efficiency and stability than the baseline agents. Based on its theoretical benefits, we believe that the NN approximator can be further applied to other complex domains to speed-up learning.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17151

49.TempLe: Learning Template of Transitions for Sample Efficient Multi-task RL

Abstract: Transferring knowledge among various environments is important for efficiently learning multiple tasks online. Most existing methods directly use the previously learned models or previously learned optimal policies to learn new tasks. However, these methods may be inefficient when the underlying models or optimal policies are substantially different across tasks. In this paper, we propose Template Learning (TempLe), a PAC-MDP method for multi-task reinforcement learning that could be applied to tasks with varying state/action space without prior knowledge of inter-task mappings. TempLe gains sample efficiency by extracting similarities of the transition dynamics across tasks even when their underlying models or optimal policies have limited commonalities. We present two algorithms for an `online'' and a`finite-model'' setting respectively. We prove that our proposed TempLe algorithms achieve much lower sample complexity than single-task learners or state-of-the-art multi-task methods. We show via systematically designed experiments that our TempLe method universally outperforms the state-of-the-art multi-task methods (PAC-MDP or not) in various settings and regimes.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17174

50.Foresee then Evaluate: Decomposing Value Estimation with Latent Future Prediction

Abstract: Value function is the central notion of Reinforcement Learning (RL). Value estimation, especially with function approximation, can be challenging since it involves the stochasticity of environmental dynamics and reward signals that can be sparse and delayed in some cases. A typical model-free RL algorithm usually estimates the values of a policy by Temporal Difference (TD) or Monte Carlo (MC) algorithms directly from rewards, without explicitly taking dynamics into consideration. In this paper, we propose Value Decomposition with Future Prediction (VDFP), providing an explicit two-step understanding of the value estimation process: 1) first foresee the latent future, 2) and then evaluate it. We analytically decompose the value function into a latent future dynamics part and a policy-independent trajectory return part, inducing a way to model latent dynamics and returns separately in value estimation. Further, we derive a practical deep RL algorithm, consisting of a convolutional model to learn compact trajectory representation from past experiences, a conditional variational auto-encoder to predict the latent future dynamics and a convex return model that evaluates trajectory representation. In experiments, we empirically demonstrate the effectiveness of our approach for both off-policy and on-policy RL in several OpenAI Gym continuous control tasks as well as a few challenging variants with delayed reward.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17182

51.Iterative Bounding MDPs: Learning Interpretable Policies via Non-Interpretable Methods

Abstract: Current work in explainable reinforcement learning generally produces policies in the form of a decision tree over the state space. Such policies can be used for formal safety verification, agent behavior prediction, and manual inspection of important features. However, existing approaches fit a decision tree after training or use a custom learning procedure which is not compatible with new learning techniques, such as those which use neural networks. To address this limitation, we propose a novel Markov Decision Process (MDP) type for learning decision tree policies: Iterative Bounding MDPs (IBMDPs). An IBMDP is constructed around a base MDP so each IBMDP policy is guaranteed to correspond to a decision tree policy for the base MDP when using a method-agnostic masking procedure. Because of this decision tree equivalence, any function approximator can be used during training, including a neural network, while yielding a decision tree policy for the base MDP. We present the required masking procedure as well as a modified value update step which allows IBMDPs to be solved using existing algorithms. We apply this procedure to produce IBMDP variants of recent reinforcement learning methods. We empirically show the benefits of our approach by solving IBMDPs to produce decision tree policies for the base MDPs.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17192

52.Toward Robust Long Range Policy Transfer

Abstract: Humans can master a new task within a few trials by drawing upon skills acquired through prior experience. To mimic this capability, hierarchical models combining primitive policies learned from prior tasks have been proposed. However, these methods fall short comparing to the human's range of transferability. We propose a method, which leverages the hierarchical structure to train the combination function and adapt the set of diverse primitive polices alternatively, to efficiently produce a range of complex behaviors on challenging new tasks. We also design two regularization terms to improve the diversity and utilization rate of the primitives in the pre-training phase. We demonstrate that our method outperforms other recent policy transfer methods by combining and adapting these reusable primitives in tasks with continuous action space. The experiment results further show that our approach provides a broader transferring range. The ablation study also show the regularization terms are critical for long range policy transfer. Finally, we show that our method consistently outperforms other methods when the quality of the primitives varies.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17196

53.Expected Eligibility Traces

Abstract: The question of how to determine which states and actions are responsible for a certain outcome is known as the credit assignment problem and remains a central research question in reinforcement learning and artificial intelligence. Eligibility traces enable efficient credit assignment to the recent sequence of states and actions experienced by the agent, but not to counterfactual sequences that could also have led to the current state. In this work, we introduce expected eligibility traces. Expected traces allow, with a single update, to update states and actions that could have preceded the current state, even if they did not do so on this occasion. We discuss when expected traces provide benefits over classic (instantaneous) traces in temporal-difference learning, and show that some- times substantial improvements can be attained. We provide a way to smoothly interpolate between instantaneous and expected traces by a mechanism similar to bootstrapping, which ensures that the resulting algorithm is a strict generalisation of TD(λ). Finally, we discuss possible extensions and connections to related ideas, such as successor features.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17200

54.Adaptive Algorithms for Multi-armed Bandit with Composite and Anonymous Feedback

Abstract: We study the multi-armed bandit (MAB) problem with composite and anonymous feedback. In this model, the reward of pulling an arm spreads over a period of time (we call this period as reward interval) and the player receives partial rewards of the action, convoluted with rewards from pulling other arms, successively. Existing results on this model require prior knowledge about the reward interval size as an input to their algorithms. In this paper, we propose adaptive algorithms for both the stochastic and the adversarial cases, without requiring any prior information about the reward interval. For the stochastic case, we prove that our algorithm guarantees a regret that matches the lower bounds (in order). For the adversarial case, we propose the first algorithm to jointly handle non-oblivious adversary and unknown reward interval size. We also conduct simulations based on real-world dataset. The results show that our algorithms outperform existing benchmarks.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17224

55.Self-correcting Q-learning

Abstract: The Q-learning algorithm is known to be affected by the maximization bias, i.e. the systematic overestimation of action values, an important issue that has recently received renewed attention. Double Q-learning has been proposed as an efficient algorithm to mitigate this bias. However, this comes at the price of an underestimation of action values, in addition to increased memory requirements and a slower convergence. In this paper, we introduce a new way to address the maximization bias in the form of a "self-correcting algorithm" for approximating the maximum of an expected value. Our method balances the overestimation of the single estimator used in conventional Q-learning and the underestimation of the double estimator used in Double Q-learning. Applying this strategy to Q-learning results in Self-correcting Q-learning. We show theoretically that this new algorithm enjoys the same convergence guarantees as Q-learning while being more accurate. Empirically, it performs better than Double Q-learning in domains with rewards of high variance, and it even attains faster convergence than Q-learning in domains with rewards of zero or low variance. These advantages transfer to a Deep Q Network implementation that we call Self-correcting DQN and which outperforms regular DQN and Double DQN on several tasks in the Atari 2600 domain.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17334

56.Self-Supervised Attention-Aware Reinforcement Learning

Abstract: Visual saliency has emerged as a major visualization tool for interpreting deep reinforcement learning (RL) agents. However, much of the existing research uses it as an analyzing tool rather than an inductive bias for policy learning. In this work, we use visual attention as an inductive bias for RL agents. We propose a novel self-supervised attention learning approach which can 1. learn to select regions of interest without explicit annotations, and 2. act as a plug for existing deep RL methods to improve the learning performance. We empirically show that the self-supervised attention-aware deep RL methods outperform the baselines in the context of both the rate of convergence and performance. Furthermore, the proposed self-supervised attention is not tied with specific policies, nor restricted to a specific scene. We posit that the proposed approach is a general self-supervised attention module for multi-task learning and transfer learning, and empirically validate the generalization ability of the proposed method. Finally, we show that our method learns meaningful object keypoints highlighting improvements both qualitatively and quantitatively.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17235

57.Physics-constrained Automatic Feature Engineering for Predictive Modeling in Materials Science

Abstract: Automatic Feature Engineering (AFE) aims to extract useful knowledge for interpretable predictions given data for the machine learning tasks. Here, we develop AFE to extract dependency relationships that can be interpreted with functional formulas to discover physics meaning or new hypotheses for the problems of interest. We focus on materials science applications, where interpretable predictive modeling may provide principled understanding of materials systems and guide new materials discovery. It is often computationally prohibitive to exhaust all the potential relationships to construct and search the whole feature space to identify interpretable and predictive features. We develop and evaluate new AFE strategies by exploring a feature generation tree (FGT) with deep Q-network (DQN) for scalable and efficient exploration policies. The developed DQN-based AFE strategies are benchmarked with the existing AFE methods on several materials science datasets.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17247

58.Domain Adaptation In Reinforcement Learning Via Latent Unified State Representation

Abstract: Despite the recent success of deep reinforcement learning (RL), domain adaptation remains an open problem. Although the generalization ability of RL agents is critical for the real-world applicability of Deep RL, zero-shot policy transfer is still a challenging problem since even minor visual changes could make the trained agent completely fail in the new task. To address this issue, we propose a two-stage RL agent that first learns a latent unified state representation (LUSR) which is consistent across multiple domains in the first stage, and then do RL training in one source domain based on LUSR in the second stage. The cross-domain consistency of LUSR allows the policy acquired from the source domain to generalize to other target domains without extra training. We first demonstrate our approach in variants of CarRacing games with customized manipulations, and then verify it in CARLA, an autonomous driving simulator with more complex and realistic visual observations. Our results show that this approach can achieve state-of-the-art domain adaptation performance in related RL tasks and outperforms prior approaches based on latent-representation based RL and image-to-image translation.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17251

59.On Convergence of Gradient Expected Sarsa(λ)

Abstract: We study the convergence of Expected Sarsa(λ) with function approximation. We show that with off-line es- timate (multi-step bootstrapping) to ExpectedSarsa(λ) is unstable for off-policy learning. Furthermore, based on convex-concave saddle-point framework, we propose a con- vergent Gradient Expected Sarsa(λ) (GES(λ)) algorithm. The theoretical analysis shows that the proposed GES(λ) converges to the optimal solution at a linear convergence rate under true gradient setting. Furthermore, we develop a Lyapunov function technique to investigate how the step- size influences finite-time performance of GES(λ). Addition- ally, such a technique of Lyapunov function can be poten- tially generalized to other gradient temporal difference algorithms. Finally, our experiments verify the effectiveness of our GES(λ). For the details of proof, please refer to https: //arxiv.org/pdf/2012.07199.pdf.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17270

60.Sample Complexity of Policy Gradient Finding Second-Order Stationary Points

Abstract: The policy-based reinforcement learning (RL) can be considered as maximization of its objective. However, due to the inherent non-concavity of its objective, the policy gradient method to a first-order stationary point (FOSP) cannot guar- antee a maximal point. A FOSP can be a minimal or even a saddle point, which is undesirable for RL. It has be found that if all the saddle points are strict, all the second-order station- ary points (SOSP) are exactly equivalent to local maxima. Instead of FOSP, we consider SOSP as the convergence criteria to characterize the sample complexity of policy gradient. Our result shows that policy gradient converges to an $(\varepsilon, \sqrt{\varepsilon\chi})$-SOSP with probability at least $1 - O(\delta)$ after the total cost of $O(\varepsilon - 9/2)$sinificantly improves the state of the art cost $O(\varepsilon - 9)$.Our analysis is based on the key idea that decomposes the parameter space Rp into three non-intersected regions: non-stationary point region, saddle point region, and local optimal region, then making a local improvement of the objective of RL in each region. This technique can be potentially generalized to extensive policy gradient methods. For the complete proof, please refer to https: //arxiv.org/pdf/2012.01491.pdf.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17271

61.WCSAC: Worst-Case Soft Actor Critic for Safety-Constrained Reinforcement Learning

Abstract: Safe exploration is regarded as a key priority area for reinforcement learning research. With separate reward and safety signals, it is natural to cast it as constrained reinforcement learning, where expected long-term costs of policies are constrained. However, it can be hazardous to set constraints on the expected safety signal without considering the tail of the distribution. For instance, in safety-critical domains, worst-case analysis is required to avoid disastrous results. We present a novel reinforcement learning algorithm called Worst-Case Soft Actor Critic, which extends the Soft Actor Critic algorithm with a safety critic to achieve risk control. More specifically, a certain level of conditional Value-at-Risk from the distribution is regarded as a safety measure to judge the constraint satisfaction, which guides the change of adaptive safety weights to achieve a trade-off between reward and safety. As a result, we can optimize policies under the premise that their worst-case performance satisfies the constraints. The empirical analysis shows that our algorithm attains better risk control compared to expectation-based methods.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17272

62.Improving Sample Efficiency in Model-Free Reinforcement Learning from Images

Abstract: Training an agent to solve control tasks directly from high-dimensional images with model-free reinforcement learning (RL) has proven difficult. A promising approach is to learn a latent representation together with the control policy. However, fitting a high-capacity encoder using a scarce reward signal is sample inefficient and leads to poor performance. Prior work has shown that auxiliary losses, such as image reconstruction, can aid efficient representation learning. However, incorporating reconstruction loss into an off-policy learning algorithm often leads to training instability. We explore the underlying reasons and identify variational autoencoders, used by previous investigations, as the cause of the divergence. Following these findings, we propose effective techniques to improve training stability. This results in a simple approach capable of matching state-of-the-art model-free and model-based algorithms on MuJoCo control tasks. Furthermore, our approach demonstrates robustness to observational noise, surpassing existing approaches in this setting. Code, results, and videos are anonymously available at https://sites.google.com/view/sac-ae/home.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17276

63.Sequential Generative Exploration Model for Partially Observable Reinforcement Learning

Abstract: Many challenging partially observable reinforcement learning problems have sparse rewards and most existing model-free algorithms struggle with such reward sparsity. In this paper, we propose a novel reward shaping approach to infer the intrinsic rewards for the agent from a sequential generative model. Specifically, the sequential generative model processes a sequence of partial observations and actions from the agent's historical transitions to compile a belief state for performing forward dynamics prediction. Then we utilize the error of the dynamics prediction task to infer the intrinsic rewards for the agent. Our proposed method is able to derive intrinsic rewards that could better reflect the agent's surprise or curiosity over its ground-truth state by taking a sequential inference procedure. Furthermore, we formulate the inference procedure for dynamics prediction as a multi-step forward prediction task, where the time abstraction that has been incorporated could effectively help to increase the expressiveness of the intrinsic reward signals. To evaluate our method, we conduct extensive experiments on challenging 3D navigation tasks in ViZDoom and DeepMind Lab. Empirical evaluation results show that our proposed exploration method could lead to significantly faster convergence than various state-of-the-art exploration approaches in the testified navigation domains.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17279

64.Exploration by Maximizing Renyi Entropy for Reward-Free RL Framework

Abstract: Exploration is essential for reinforcement learning (RL). To face the challenges of exploration, we consider a reward-free RL framework that completely separates exploration from exploitation and brings new challenges for exploration algorithms. In the exploration phase, the agent learns an exploratory policy by interacting with a reward-free environment and collects a dataset of transitions by executing the policy. In the planning phase, the agent computes a good policy for any reward function based on the dataset without further interacting with the environment. This framework is suitable for the meta RL setting where there are many reward functions of interest. In the exploration phase, we propose to maximize the Renyi entropy over the state-action space and justify this objective theoretically. The success of using Renyi entropy

as the objective results from its encouragement to explore the hard-to-reach state-actions. We further deduce a policy gradient formulation for this objective and design a practical exploration algorithm that can deal with complex environments. In the planning phase, we solve for good policies given arbitrary reward functions using a batch RL algorithm. Empirically, we show that our exploration algorithm is effective and sample efficient, and results in superior policies for arbitrary reward functions in the planning phase.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17297

65.Sample Efficient Reinforcement Learning with REINFORCE

Abstract: Policy gradient methods are among the most effective methods for large-scale reinforcement learning, and their empirical success has prompted several works that develop the foundation of their global convergence theory. However, prior works have either required exact gradients or state-action visitation measure based mini-batch stochastic gradients with a diverging batch size, which limit their applicability in practical scenarios. In this paper, we consider classical policy gradient methods that compute an approximate gradient with a single trajectory or a fixed size mini-batch of trajectories under soft-max parametrization and log-barrier regularization, along with the widely-used REINFORCE gradient estimation procedure. By controlling the number of "bad" episodes and resorting to the classical doubling trick, we establish an anytime sub-linear high probability regret bound as well as almost sure global convergence of the average regret with an asymptotically sub-linear rate. These provide the first set of global convergence and sample efficiency results for the well-known REINFORCE algorithm and contribute to a better understanding of its performance in practice.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17300

66.Mean-Variance Policy Iteration for Risk-Averse Reinforcement Learning

Abstract: We present a mean-variance policy iteration (MVPI) framework for risk-averse control in a discounted infinite horizon MDP optimizing the variance of a per-step reward random variable. MVPI enjoys great flexibility in that any policy evaluation method and risk-neutral control method can be dropped in for risk-averse control off the shelf, in both on- and off-policy settings. This flexibility reduces the gap between risk-neutral control and risk-averse control and is achieved by working on a novel augmented MDP directly. We propose risk-averse TD3 as an example instantiating MVPI, which outperforms vanilla TD3 and many previous risk-averse control methods in challenging Mujoco robot simulation tasks under a risk-aware performance metric. This risk-averse TD3 is the first to introduce deterministic policies and off-policy learning into risk-averse reinforcement learning, both of which are key to the performance boost we show in Mujoco domains.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17302

67.The Sample Complexity of Teaching by Reinforcement on Q-Learning

Abstract: We study the sample complexity of teaching, termed as ``teaching dimension" (TDim) in the literature, for the teaching-by-reinforcement paradigm, where the teacher guides the student through rewards. This is distinct from the teaching-by-demonstration paradigm motivated by robotics applications, where the teacher teaches by providing demonstrations of state/action trajectories. The teaching-by-reinforcement paradigm applies to a wider range of real-world settings where a demonstration is inconvenient, but has not been studied systematically. In this paper, we focus on a specific family of reinforcement learning algorithms, Q-learning, and characterize the TDim under different teachers with varying control power over the environment, and present matching optimal teaching algorithms. Our TDim results provide the minimum number of samples needed for reinforcement learning, and we discuss their connections to standard PAC-style RL sample complexity and teaching-by-demonstration sample complexity results. Our teaching algorithms have the potential to speed up RL agent learning in applications where a helpful teacher is available.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17306

68.Augmenting Policy Learning with Routines Discovered from a Single Demonstration

Abstract: Humans can abstract prior knowledge from very little data and use it to boost skill learning. In this paper, we propose routine-augmented policy learning (RAPL), which discovers routines composed of primitive actions from a single demonstration and uses discovered routines to augment policy learning. To discover routines from the demonstration, we first abstract routine candidates by identifying grammar over the demonstrated action trajectory. Then, the best routines measured by length and frequency are selected to form a routine library. We propose to learn policy simultaneously at primitive-level and routine-level with discovered routines, leveraging the temporal structure of routines. Our approach enables imitating expert behavior at multiple temporal scales for imitation learning and promotes reinforcement learning exploration. Extensive experiments on Atari games demonstrate that RAPL improves the state-of-the-art imitation learning method SQIL and reinforcement learning method A2C. Further, we show that discovered routines can generalize to unseen levels and difficulties on the CoinRun benchmark.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17316

69.Inverse Reinforcement Learning with Natural Language Goals

Abstract: Humans generally use natural language to communicate task requirements to each other. Ideally, natural language should also be usable for communicating goals to autonomous machines (e.g., robots) to minimize friction in task specification. However, understanding and mapping natural language goals to sequences of states and actions is challenging. Specifically, existing work along these lines has encountered difficulty in generalizing learned policies to new natural language goals and environments. In this paper, we propose a novel adversarial inverse reinforcement learning algorithm to learn a language-conditioned policy and reward function. To improve generalization of the learned policy and reward function, we use a variational goal generator to relabel trajectories and sample diverse goals during training. Our algorithm outperforms multiple baselines by a large margin on a vision-based natural language instruction following dataset (Room-2-Room), demonstrating a promising advance in enabling the use of natural language instructions in specifying agent goals.

HomePage: https://ojs.aaai.org/index.php/AAAI/article/view/17326
