Highlights

A Review of Safe Reinforcement Learning: Methods, Theories and Applications

Shangding Gu[a*], Long Yang[b], Yali Du[c], Guang Chen[d], Florian Walter[a], Jun Wang[e], Alois Knoll[a]

- A review of safe Reinforcement Learning (RL) methods is provided with theoretical and application analyses.
- The key question that safe RL needs to answer is proposed, and five problems **"2H3W"** are analyzed to address the key question.
- To examine the effectiveness of safe RL methods, several safe single-agent and multi-agent RL benchmarks are investigated.
- The challenging problems are pointed out to guide the research directions.

A Review of Safe Reinforcement Learning: Methods, Theories and Applications

Shangding Gu[a*], Long Yang[b], Yali Du[c], Guang Chen[d], Florian Walter[a], Jun Wang[e], Alois Knoll[a]

[a]*Department of Computer Science, Technical University of Munich, Germany*
[b]*Institute for AI, Peking University & BIGAI, China*
[c]*Department of Informatics, King's College London, UK*
[d]*College of Electronic and Information Engineering, Tongji University, China*
[e]*Department of Computer Science, University College London, UK*

Abstract

Reinforcement Learning (RL) has achieved tremendous success in many complex decision-making tasks. However, safety concerns are raised during deploying RL in real-world applications, leading to a growing demand for safe RL algorithms, such as in autonomous driving and robotics scenarios. While safe control has a long history, the study of safe RL algorithms is still in the early stages. To establish a good foundation for future safe RL research, in this paper, we provide a review of safe RL from the perspectives of methods, theories, and applications. Firstly, we review the progress of safe RL from five dimensions and come up with five crucial problems for safe RL being deployed in real-world applications, coined as **"2H3W"**. Secondly, we analyze the algorithm and theory progress from the perspectives of answering the **"2H3W"** problems. Particularly, the sample complexity of safe RL algorithms is reviewed and discussed, followed by an introduction to the applications and benchmarks of safe RL algorithms. Finally, we open the discussion of the challenging problems in safe RL, hoping to inspire future research on this thread. To advance the study of safe RL algorithms, we release an open-sourced repository containing the implementations of major safe RL algorithms at the link[1].

*This manuscript is under actively development. We appreciate any constructive comments and suggestions corresponding to *shangding.gu@tum.de*.

[1] https://github.com/chauncygu/Safe-Reinforcement-Learning-Baselines.git

Keywords: safe reinforcement learning; safety optimisation; constrained Markov decision processes; safety problems

1. Introduction

Over the past decades, Reinforcement Learning (RL) has been widely adopted in many fields, e.g. transportation schedule [25, 58, 184, 289], traffic signal control [54, 65], energy management [210], wireless security [199], satellite docking [87], edge computing [339], chemical processes [276], video games [31, 178, 222, 245, 299, 341], board games of Go, shogi, chess and arcade game PAC-MAN [146, 286, 287, 288], finance [5, 170, 302, 318], autonomous driving [116, 145, 157, 203, 219, 271], recommender systems [60, 283, 362], resource allocation [149, 194, 211], communication and networks [43, 44, 137, 192], smart grids [167, 277], video compression [338, 205], and robotics [6, 16, 39, 49, 118, 123, 140, 162, 182, 217, 234, 246, 250, 263], etc. However, a challenging problem in this domain is: **how do we guarantee safety when we deploy RL in real-world applications?** After all, unacceptable catastrophes may arise if we fail to take safety into account during RL applications in real-world scenarios. For example, it must not hurt humans when robots interact with humans in human-machine interaction environments; false or racially discriminating information should not be recommended for people in recommender systems; safety has to be ensured when self-driving cars are carrying out tasks in real-world environments. More specifically, we introduce several safety definitions from different perspectives, which might be helpful for safe RL research.

Safety definition. The first type of safety definition: according to the definition of Oxford dictionary [296], the phrase "safety" is commonly interpreted to mean "the condition of being protected from or unlikely to cause danger, risk, or injury." The second type of safety definition: the definition of general "safety" according to wiki [2], the state of being "safe" is defined as "being protected from harm or other dangers"; "controlling recognized dangers to attain an acceptable level of risk" is also referred to as "safety". The third type of safety definition: according to Hans *et al.* [128], humans need to label environmental states as "safe" or "unsafe," and agents are considered "safe" if "they never reach unsafe states". The fourth type of safety definition:

[2] https://en.wikipedia.org/wiki/Safety

agents are considered to be "safe" by some research [126, 144, 180] if "they act, reason, and generalize obeying human desires". The fifth type of safety definitions: Moldovan and Abbeel [224] consider an agent "safe" if "it meets an ergodicity requirement: it can reach each state it visits from any other state it visits, allowing for reversible errors". In this review, based on the above various definitions, we investigate safe RL methods, which are about optimizing cost objectives, avoiding adversary attacks, improving undesirable situations, reducing risk, and controlling agents to be safe, etc.

RL safety is a significant practical problem that confronts us in RL applications, and is one of the critical problems in AI safety [14] that remains unsolved, though it has attracted increasing attention in the field of RL. Moreover, it is deduced that the mean minus variance [206] and percentile optimisation [78] of safe RL are, in general, NP-hard problems [223]. In some applications, the agent's safety is much more important than the agent's reward [86]. An attempt to answer the question above raises some fundamental problems that we call **"2H3W"** problems:

(1) **Safety Policy. H**ow can we perform policy optimisation to search for a safe policy?

(2) **Safety Complexity. H**ow much training data is required to find a safe policy?

(3) **Safety Applications. W**hat is the up to date progress of safe RL applications?

(4) **Safety Benchmarks. W**hat benchmarks can we use to fairly and holistically examine safe RL performance?

(5) **Safety Challenges. W**hat are the challenges faced in future safe RL research?

Most of the research in this field is aimed at solving the above **"2H3W"** problems, and the framework of safe RL about **"2H3W"** problems is shown in Figure 1.

As for the problem (1) (**safety policy**), in many practical applications, a robot must not visit some states, and must not take some actions, which can be thought of as "unsafe" either for itself or for elements of its environment. It is essential that a safe policy function or value function is provided so

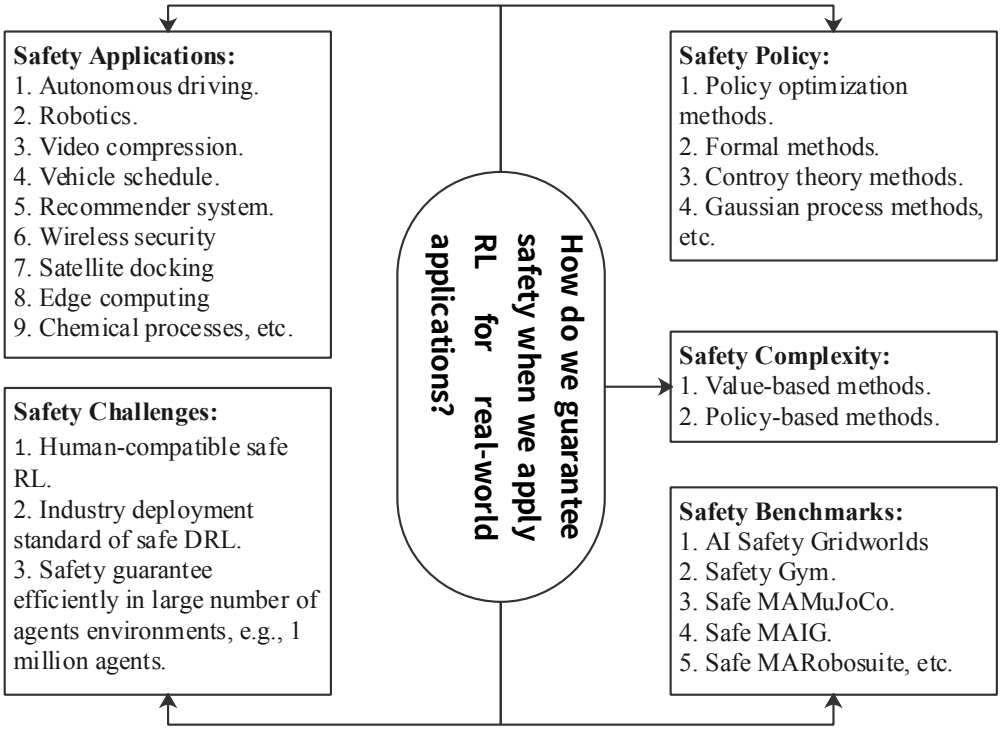


Figure 1: The framework of safe RL about **"2H3W"** problems.

that agents can reach safe states or perform safe actions. To achieve safety policy, a growing number of approaches have been developed over the last few decades, such as primal-dual policy optimization methods [55, 177, 186, 241, 270, 274, 310, 342], trust region policy optimization with safety constraints methods [6, 35, 37, 127, 324, 348], formal methods [15, 26, 28, 101, 102, 143, 219, 242, 264, 317, 321, 329, 366], control theory-based methods [59, 62, 63, 83, 142, 147, 173, 247, 248], Gaussian processes methods [8, 51, 70, 94, 253, 276, 297, 319, 326].

As for problem (2) (**safety complexity**), agents need to interact with environments and sample trajectories, such that the safe RL algorithms converge to below the constraint bound, and guarantee application safety. Since one of the natures of RL is exploration learning [300], it is usually hard to control the balance between exploitation and exploration, especially when we need to improve reward performance while satisfying cost constraints (cost is one way of encoding safety). Thus, we need to determine how many sample trajectories can make safe RL algorithms converge and satisfy safety

bounds using the analysis of safety complexity during safe RL applications. Furthermore, the sample complexity of each safe RL algorithm is investigated from the viewpoints of value-based [81, 131, 151, 221], and policy-based [82, 343, 357] methods, etc.

As for problem (3) (**safety applications**), although there are many RL applications to date, most of the applications are merely simulations that do not take safety into account; some real-world experiments have been carried out, but there is still a long way to go before RL can be used in real-world applications. Generally, when we use safe RL in real-world applications, we need to consider the ego agent safety, environmental safety, and human safety. Most importantly, we need to consider the control safety, which prevents adversary attacks from destroying or controlling the agent [133]. Therefore, for the safe RL application research, we introduce some safe RL applications in this review, e.g., autonomous driving [116, 145, 169, 154, 157, 218, 219, 328], robotics [106, 108, 248, 250, 292, 311], video compression [205], vehicle schedule [25, 184], etc.

As for problem (4) (**safety benchmarks**), we need to determine how to design cost and reward functions considering the balance between RL reward performance and safety in each benchmark, since cost functions will typically disturb reward performance. If we have a loose cost function, we may not be able to guarantee agent safety during the learning process; if we take too conservative cost functions, for example, in a constrained policy optimization process, when we set the cost constraint bound as zero or a negative value, which may result in lousy reward performance. Thus, we should pay more attention to designing the cost and reward function in the benchmarks. In some safe RL benchmarks, such as AI Safety Gridworlds [181], Safety Gym [259], Safe MAMuJoCo [118], Safe MAIG [3], Safe MARobosuite [118], the cost and reward functions are tailored to specific tasks well in examining experiments.

As for problem (5) (**safety challenges**), firstly, when we consider RL safety, it is a significant challenge about how to consider human safety factors or environmental safety factors during deploying RL in real-world applications. Secondly, a further important aspect when considering RL safety is how to take robot safety factors into account during RL applications [14, 133]. Another critical challenge is the social dilemma problem with safety balance. For

[3] https://github.com/chauncygu/Safe-Multi-Agent-Isaac-Gym.git

example, the game of the trolley problem [314]. In the game, an out-of-control trolley will eventually kill five people if no action is taken, but you can redirect the trolley to another track, where only one person will be killed. The open question is how to balance safety wights when using RL. Moreover, the application standard and safe Multi-Agent RL (MARL) should be considered for future research.

The problem (5) (**safety challenges**) appear to be dilemma problems. They are not more straightforward problems compared to problem (1) (**safety policy**), problem (2) (**safety complexity**), problem (3) (**safety applications**), and the problem (4)(**safety benchmarks**). We must guarantee agent, human, and environmental safety when we apply RL for practical applications by providing sophisticated algorithms. In addition, few studies have focused on problem (2) (**safety complexity**) and problem (4) (**safety benchmarks**), especially in industrial use. Answering problem (3) (**safety applications**) and problem (5) (**safety challenges**) may reveal RL application situations and provide a clue for future RL research. In this review, we will summarise the progress of safe RL, answer the five problems and analyze safe RL algorithms, theory, and applications. In general, we mainly sort out the safe RL research of the past two decades (Though we are unfortunately unable to include some impressive safe RL literature in this review for space reasons).

The main contributions of this paper: first, we investigate safe RL research and give an indication of the research progress. Second, the main practical question of RL applications is discussed, and five fundamental problems are analyzed in detail. Third, algorithms, theory, and applications of safe RL are reviewed in detail, e.g., safe model-based learning and safe model-free learning, in which we present a bird's eye view to summarising the progress of safe RL. Finally, the challenges we face when using RL for applications are explained.

The remainder of this paper is organized as follows: safe RL background, including safe RL problem formulation and related safe RL surveys, are introduced in Section 2; an overview of safe RL is provided in Section 3; safe RL theory and algorithm analysis are presented in Section 4; the application analysis of safe RL is introduced in Section 5; several safe RL benchmarks are introduced in detail in Section 6; challenging problems and remaining questions are described in Section 7; a conclusion for the review is given in Section 8.

2. Background

Safe reinforcement learning is often modeled as a Constrained Markov Decision Process (CMDP) [12], in which we need to maximize the agent reward while making agents satisfy safety constraints. A substantial body of literature has studied Constrained Markov Decision Process (CMDP) problems for both tabular and linear cases [12, 33, 34, 152, 267, 268, 269]. However, deep safe RL for high dimensional and continuous CMDP optimization problems is a relatively new area that has emerged in recent years, and proximal optimal values generally represent safe states or actions using neural networks. In this section, we illustrate the generally deep safe RL problem formulation concerning the objective functions of safe RL and offer an introduction to safe RL surveys.

2.1. Problem Formulation of Safe Reinforcement Learning

A CMDP problem [12] is an extension of a standard Markov decision process (MDP) $\mathcal{M}$ with a constraint set $\mathcal{C}$. A tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \rho_0, \gamma)$ is given to present a MDP [255]. A state set is denoted as $\mathcal{S}$, an action set is denoted as $\mathcal{A}$, $\mathbb{P}(s^{'}|s, a)$ denotes the probability of state transition from s to s' after playing a. A reward function is denoted as $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. $\rho_0(\cdot) : \mathcal{S} \to [0, 1]$ is the starting state distribution, γ denotes the discount factor.

In safe RL, the goal of an optimal policy π is to maximize the reward and minimize the cost by selecting an action a, and $\mathbf{\Pi_S}$ is the policy set, τ is a trajectory, $\tau = (s_0, a_0, s_1, \ldots)$, in which an action depends on π, $s_0 \sim \rho_0(\cdot)$, $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$. $d_\pi^{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_\pi(s_t = s|s_0)$ denotes the state distribution (starting at s_0), the discounted state distribution based on the initial distribution $\rho_0(\cdot)$ is present as $d_\pi^{\rho_0}(s) = \mathbb{E}_{s_0\sim\rho_0(\cdot)}[d_\pi^{s_0}(s)]$.

The state value function is shown in Function (1), the state action value function is shown in Function (2); the advantage function is shown in Function (3); the reward objective is shown in Function (4).

$$V_\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r_{t+1}|s_0 = s]. \tag{1}$$

$$Q_\pi(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty}\gamma^t r_{t+1}|s_0 = s, a_0 = a]. \tag{2}$$

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s). \tag{3}$$

$$J(\pi) = \mathbb{E}_{s\sim\rho_0(\cdot)}[V_\pi(s)]. \tag{4}$$

CMDP is based on MDP, in which the additional constraint set $C = \{(c_i, b_i)\}_{i=1}^m$ is considered, where c_i is the cost value functions, and b_i is the safety constraint bound, $i \in [m]$, m is the type number of cost constraints. Similarly, we have the cost value functions $V_\pi^{c_i}$, cost action-value functions $Q_\pi^{c_i}$, and cost advantage functions $A_\pi^{c_i}$, e.g., $V_\pi^{c_i}(s) = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t c_i(s_t, a_t)|s_0 = s\right]$. Based on the above definition, we have the expected cost function $C_i(\pi) = \mathbb{E}_{s\sim\rho_0(\cdot)}[V_\pi^{c_i}(s)]$. The feasible policy set Π_C that can satisfy the safety constraint bound is as follows,

$$\Pi_C = \cap_{i=1}^m \left\{\pi \in \Pi_{\mathrm{S}} \quad \text{and} \quad C_i(\pi) \le b_i\right\}.$$

The goal of safe RL is to maximise the reward performance and minimise the cost values to guarantee safety,

$$\max_{\pi\in\Pi_{\mathrm{S}}} J(\pi), \text{ such that } \mathbf{c}(\pi) \preceq \mathbf{b}, \tag{5}$$

where the vector $\mathbf{c}(\pi) = (C_1(\pi), C_2(\pi), \cdots, C_m(\pi))^\top$, and $\mathbf{b} = (b_1, b_2, \cdots, b_m)^\top$. Therefore, the solution of a CMDP problem (5) can be provided as the optimization function (6):

$$\pi_\star = \arg\max_{\pi\in\Pi_C} J(\pi). \tag{6}$$

2.2. A Survey of Safe Reinforcement Learning related surveys

Several surveys have already investigated safe RL problems and methods, e.g., [49, 107, 159, 195]. However, the surveys do not provide comprehensive theoretical analysis for safe RL, such as sample complexity, nor do they focus on the critical problems of safe RL that we point out, **"2H3W"** problems. For example, [49] investigates safe RL from the perspective of control theory and robotics; the soft constraints, probabilistic constraints, and hard constraints are defined in the the survey. Furthermore, they reviewed a large number of papers that are related to control theory and analyzed how to use control

theory to guarantee RL safety and stability, such as using model predictive control [17, 156, 164, 213, 236, 266, 295], adaptive control [104, 138, 231, 273], robust control [29, 84, 135, 322], Lyapunov functions [62, 63] for RL stability and safety. In the survey [107], they focus more on reviewing safe RL methods up to 2015. They categorize safe RL methods into two types: one is based on the safety of optimization criterion, where the worst case, risk-sensitive criterion, constrained criterion, etc., are taken into account to ensure safety. Another one is based on external knowledge or risk metric. In general, the external knowledge or risk metric is leveraged to guide the optimisation of RL safety. In contrast to the survey [107], the survey [159] focuses more on the techniques of safe learning, including MDP and non-MDP methods, such as RL, active learning, evolutionary learning. The survey [195] summarises safe model-free RL methods based on two kinds of constraints, namely cumulative constraints and instantaneous constraints.

As for safe RL methods of cumulative constraints, three types of cumulative constraints are introduced:
a discounted cumulative constraint:

$$J_{C_i}^{\pi_\theta} = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t C_i\left(s_t, a_t, s_{t+1}\right)\right] \leq b_i, \tag{7}$$

a mean valued constraint:

$$J_{C_i}^{\pi_\theta} = \mathbb{E}_{\tau\sim\pi_\theta}\left[\frac{1}{T}\sum_{t=0}^{T-1} C_i\left(s_t, a_t, s_{t+1}\right)\right] \leq b_i, \tag{8}$$

a probabilistic constraint:

$$J_{C_i}^{\pi_\theta} = P\left(\sum_{t} C_i\left(s_t, a_t, s_{t+1}\right) \leq b_i\right) \geq \eta. \tag{9}$$

When it comes to instantaneous constraints, the instantaneous explicit constraints and instantaneous implicit constraints are given. The explicit ones have an accurate, closed-form expression that can be numerically checked, e.g., the cost that an agent generates during each step; the implicit does not have an accurate closed-form expression, e.g., the probability that an agent will crash into unsafe areas during each step. Although based on our investigation, most CMDP methods are based on cumulative cost optimization, a few CMDP

methods focus on the immediate costs to optimize performance [261], and it is natural to take the cost of a whole trajectory rather than a state or action in some real-world applications, such as robot-motion planning and resource allocation [62]. Compared to the related surveys [49, 107, 159, 195], our survey pays more attention to answering the **"2H3W"** problems and provides safe RL algorithm analysis, sampling complexity analysis and convergence investigation from the perspectives of model-based and model-free RL.

3. Methods of Safe Reinforcement Learning

There are several types of safe RL methods based on different criteria. For example, from the perspective of objective optimization, several methods consider cost as one of the optimization objectives to achieve safety, e.g., [27, 40, 132, 139, 150, 200, 233, 275, 302]. From the perspective of knowledge utilization, some methods consider safety in the RL exploration process by leveraging external knowledge, e.g., [4, 62, 67, 109, 224, 305, 312]. From the perspectives of policy and value-based methods for safe RL, summarise as follows, policy-based safe RL: [6, 158, 82, 118, 343, 348, 357], value-based safe RL: [23, 81, 89, 131, 151, 191, 221, 333]. From the perspective of the agent number, we have safe RL methods in a safe single-agent RL setting and a safe multi-agent RL setting. More specifically, numerous safe RL methods are about the single-agent setting. An agent needs to explore the environments to improve its reward while keeping costs below the constraint bounds. In contrast to safe single-agent RL methods, safe multi-agent RL methods not only need to consider the ego agent's reward and other agent's reward, but also have to take into account the ego agent's safety and other agents' safety in an unstable multi-agent system.

In this section, we provide a concise but holistic overview of safe RL methods from a bird's eye view and attempt to answer the **"Safety Policy"** problem. Especially, we will introduce safe RL methods from the perspectives of model-based methods and model-free methods, in which safe single-agent RL and multi-agent RL will be analyzed in detail. In particular, we will introduce policy optimization-based approaches, formal methods-based approaches, control theory-based approaches, and Gaussian processes-based approaches in model-based and model-free settings, respectively.

In addition, the model-based and model-free safe RL analysis are summarised in Table 1 and Table 2 respectively.

3.1. Methods of Model-Based Safe Reinforcement Learning

Although accurate models are challenging to build and many applications lack models, model-based Deep RL (DRL) methods usually have a better learning efficiency than model-free DRL methods. There are still many scenarios for which we can apply model-based DRL methods, such as robotics, transportation planning, logistics, etc. Several works have shown that safe problems, such as a safe robot control problem [309], can be overcome by using model-based safe RL methods. In this section, we will investigate model-based safe RL methods from different perspectives: policy optimization-based approaches, control theory-based approaches, formal methods-based approaches, and Gaussian process-based approaches.

3.1.1. Policy optimization-based approaches

Policy optimization-based approaches usually search for a safe policy using cumulative cost values on trajectories. For example, Moldovan and Abbeel use the Chernoff function (10) in [223] to achieve near-optimal bounds with desirable theoretical properties. The cumulative expectation cost is represented in the function as $E_{s,\pi}\left[e^{J/\theta}\right]$, especially for δ, which can be used to adjust the balance of reward performance and safety. For instance, if δ is set as 1, this method will ignore the safety, and if δ is set as 0, this method will fully consider the risk and optimize the cost. Moreover, they examine the method in grid world environments and air travel planning applications. However, the method needs a significant amount of time to recover policy from risk areas, which is ten times the value iteration.

$$C^{\delta}_{s,\pi}[J] = \inf_{\theta>0}\left(\theta \log E_{s,\pi}\left[e^{J/\theta}\right] - \theta \log(\delta)\right) \tag{10}$$

Different from the Chernoff function based methods [223], Borkar [41] proposes an actor-critic RL method to handle a CMDP problem based on the envelope theorem in mathematical economics [209, 216], in which the primal-dual optimization is analyzed in detail using a three-time scale process. The critic scheme, actor scheme, and dual ascent are on the fast, middle, and slow timescale. Bharadhwaj *et al.* [35] also present an actor-critic method to address a safe RL problem, where they first develop a conservative safety critic to estimate the safety. The primal-dual gradient descent is leveraged to optimize the reward and cost value by constraining the failure probability. Although this method can bound the probability of failures during policy

evaluation and improvement, this method still cannot guarantee total safety; a few unsafe corner actions may dramatically damage critical robot applications.

Akin to Borkar's method [41], Tessler *et al.* [310] also utilize a multi-timescale approach with regards to cost as a part of the reward in primal-dual methods. However, the method's learning rate is hard to tune for real-world applications because of imposing stringent requirements, and the method may not guarantee safety when agents are training.

Similar to the above methods, in actor-critic settings, with primal-dual optimization techniques, Yu *et al.*[352] convert a non-convex constrained problem into a locally convex problem and guarantee the stationary point to the optimal point of non-convex optimization problem; they consider the state-action safety optimization, and the optimization process is motivated by [188], in which the Lipschitz condition is necessary to satisfy the results. Also, they need to estimate the policy gradient for optimization.

Analogously, using optimization theory, the policy gradient and actor-critic methods are proposed by Chow *et al.* [60] to optimize risk RL performance, in which CVaR[4]-constrained and chance-constrained optimization are used to guarantee safety. Specifically, the importance sampling [24, 303] is used to improve policy estimation and provide convergence proof for the proposed algorithms. Nevertheless, this method may not guarantee safety during training [361]. Paternain *et al.* [241] provide a duality theory for CMDP optimization, and they prove the zero duality gap in primal-dual optimization even for non-convexity problems. Furthermore, they point out that the primal problem can be exactly solved by dual optimization. In their study, the suboptimal bound using neural network parametrization policy is also present [136].

3.1.2. Control theory-based approaches

Control theory-based approaches mostly ensure more rigorous safety than policy optimization approaches. Nevertheless, control theory-based approaches may not be easy to transfer to other domains due to the strict requirements of dynamics models. For instance, Berkenkamp *et al.* [30] develop a safe model-based RL algorithm by leveraging Lyapunov functions to guarantee stability with the assumptions of Gaussian process prior; their method can ensure a high-probability safe policy for an agent in a continuous

[4]CVaR denotes the Conditional Value at Risk.

process. However, Lyapunov functions are usually hand-crafted, and it is not easy to find a principle to construct Lyapunov functions for an agent's safety and performance [62]. In addition, some safe RL methods are proposed from the perspective of Model Predictive Control (MPC) [356], e.g., MPC is used to make robust decisions in CMDPs [17] by leveraging a constrained MPC method [212], which also introduces a general safety framework to make decisions [98].

Apart from Lyapunov functions and MPC-safe RL approaches, control barrier functions (CBFs) based approaches are also proposed to ensure learning safety [202, 59, 93, 68]. Similar to the above control methods, CBFs based safe RL approaches also require a dynamics model for the control system. Thus, it is not straightforward to deploy it in RL [208]. For example, [202] propose a model-based safe RL by leveraging lagrangian optimization and CBFs; they further deploy their method in simulation and real-world autonomous driving experiments, the experiment results indicate that their method can improve sample efficiency and ensure safety. In contrast to [202], [93] not just use CBFs, but also robust CBFs [92], they develop a safe RL method based on SAC [124] and robust CBFs. Specifically, they consider safety in robust settings by estimating the disturbed dynamics systems with GP models. [68] propose a safe exploration framework based on model-based RL and CBFs, where Lyapunov-like CBFs [238] are used to construct the safety CBFs, and they can ensure the learning safety and stability. The above methods show impressive performance in ensuring safety. However, designing CBFs necessitates knowledge of the model or safety certificates, which may be challenging in complex environments, even when the model is derived using alternative methods.

3.1.3. Formal methods-based approaches

Different from policy optimization-based approaches, control theory-based approaches, and Gaussian Process-based methods, formal methods [15] usually try to ensure safety without unsafe probabilities. However, most formal methods rely heavily on the model knowledge and may not show better reward performance than other methods. The verification computation might be expensive for each neural network [15]. More generally, the curse of dimensionality problem is challenging to be solved, which appears when formal methods are deployed for RL safety [28], since formal methods may be intractable to verify RL safety in continuous and high-dimension space settings [28].

For instance, in [15], Anderson *et al.* provide a neurosymbolic RL method by leveraging formal verification to guarantee RL safety in a mirror descent framework. In their method, the neurosymbolic and constrained classes using symbolic policies are leveraged to approximate gradient and conduct verification in a loop-iteration setting. The experiment results show promising performance in RL safety, though the fixed worst-case model knowledge is used in these environments, which may not be suitable for practical applications. Similarly, in [28], Beard and Baheri utilize formal methods to improve agent safety by incorporating external knowledge and penalizing behaviors for RL exploration. Nonetheless, the methods may need to be developed for scalability and continuous systems. Finally, in [101], Fulton and Platzer also use external knowledge to ensure agent safety by leveraging the justified speculative control sandbox in offline formal verification settings.

3.1.4. Gaussian processes-based approaches

In formal methods, determining how to measure unsafe areas is a challenging problem. Recently, many approaches have been proposed by leveraging a Gaussian Process (GP) [335] to estimate the uncertainty and unsafe areas. Further, the information from GP methods is incorporated into the learning process for agent safety. For instance, Akametalu *et al.* [8] develop a safe RL method based on reachability analysis, in which they use GP methods to measure the disturbances which may lead to unsafe states for agents, the maximal safe areas are computed iteratively in an unknown dynamics system. Like Akametalu *et al.* [8] method, Berkenkamp and Schoellig [29] utilize GP methods to measure the system uncertainty and further guarantee system stability. Polymenakos *et al.* [253]develop a safe policy search approach based on PILCO (Probabilistic Inference for Learning Control) method [77] which is a policy gradient method derived from a GP, in which they improve agent safety using probability trajectory predictions by incorporating cost into reward functions. Similar to Polymenakos *et al.* [253], Cowen *et al.* [70] also use PILCO to actively explore environments while considering risk, a GP is used to quantify the uncertainty during exploration, and agent safety probability is improved by leveraging a policy multi-gradient solver.

The above safe RL methods present excellent performance in terms of the balance between reward and safety performance in most challenging tasks. Nonetheless, training safety or stability may need to be further investigated rigorously, and a unified framework may need to be proposed to better examine safe RL performance.

Model-Based Safe RL	Features	Methods	Convergence Analysis
[41, Borkar (2005)]	Based on the Lagrange multiplier formulation and the saddle point property.	An actor-critic algorithm for CMDP.	YES.
[60, Chow et al., (2018)]	Two risk-constrained MDP problems. The first one involves a CVaR constraint and the second one involves a chance constraint.	Policy gradient and actor-critic algorithms via CVaR.	YES.
[310, Tessler et al., (2019)]	A novel constrained actor-critic approach RCPO[5] uses a multi-timescale approach.	Reward constrained policy optimization.	YES.
[352, Yu et al., (2019)]	Successive convex relaxation algorithm for CMDP.	Actor-Critic update for constrained MDP.	YES.
[166, Koppel et al., (2019)]	Saddle points with a reproducing Kernel Hilbert Space.	Projected Stochastic Primal-Dual Method for CMDP.	YES.
[253, Polymenakos et al., (2019)]	Ensure safety during training with high probability based on a PILCO framework.	Policy gradient with a GP model.	NO.
[241, Paternain et al., (2019).]	Solving non-convex CMDP.	Primal-dual algorithms for constrained reinforcement learning.	YES.
[35, Bharadhwaj et al., (2021)]	Without strong assumptions, e.g., unsafe state knowledge is required a prior, the method can avoid unsafe actions during training with high probability.	Conservative critics for safety exploration.	YES.
[221, Miryoosefi and Jin (2021)]	CMDP with general convex constraints.	Reward-free approach to constrained reinforcement learning.	NO.
[240, Paternain et al., (2022)]	Safety into the problem as a probabilistic version of positive invariance.	Stochastic Primal-Dual for Safe Policies.	NO.

Table 1: Model-based safe RL analysis.

3.2. Methods of Model-Free Safe Reinforcement Learning

Most studies pay attention to model-free safe RL since it can be deployed in many domains without requiring model dynamics. In this section, we also investigate safe RL methods from the perspectives of policy optimization, control theory, Gaussian processes, and formal methods.

3.2.1. Policy optimization-based approaches

Constrained Policy Optimisation (CPO)[6] is the first policy gradient method to solve the CMDP problem based on model-free deep RL. In particular, a policy has to be optimized to guarantee the reward of a monotonic improvement while satisfying safety constraints. As a result, their methods can almost converge to safety bound and produce more comparable performance than the primal-dual method [60] on some tasks. However, CPO's computation is more expensive than PPO[6]-Lagrangian, since it needs to compute the Fisher information matrix and uses the second Taylor expansion to optimize objectives. Moreover, the approximation and sampling errors may have detrimental effects on the overall performance, and the convergence analysis is challenging. Furthermore, the additional recovery policy may require more samples, which could result in wasted samples [361].

Derived from CPO [6], Projection-based Constrained Policy Optimisation (PCPO) [348] based on two-step methods constructs a cost projection to optimize cost and guarantee safety, which displays better performance than CPO on some tasks. PCPO leverages policy to maximize the reward via Trust Region Policy Optimization (TRPO) method [279], and then projects the policy to a feasible region to satisfy safety constraints. However, the second-order proximal optimization is used in both steps, which may result in a more expensive computation than the First Order Constrained Optimization in Policy Space (FOCOPS) method [361], which only uses the first-order optimization. [361] is motivated by the optimization-based idea [249], where they use the primal-dual method, address policy search in the nonparametric space and project the policy into the parameter space, to carry out proximal maximization optimization in CMDPs. Although this method is easy to implement and shows better sample efficiency, it still needs to solve the problems of unstable saddle points and unsafe actions during training.

Similar to CPO [6], in Safe Advantage-based Intervention for Learning

[6]PPO denotes the Policy Proximal Optimization.

policies with Reinforcement (SAILR) [327], Wagener *et al.* leverage the advantage function as a surrogate to minimize cost, and further achieve safe policy both during training and deployment. Furthermore, In [294], a Shortest-Path Reinforcement Learning (SPRL) method is proposed using off-policy strategies to construct safe policy and reward policy, and its applications are used for the shortest path problems in Travel Sale Path (TSP). Besides, based on a Gaussian process of a safe RL method [326], the SNO-MDP[7] [325] is developed to optimize cost in the safe region and optimize the reward in the unknown safety-constrained region [297, 319]. It is suggested that the maximization reward is more important than safety. The policy is often substantial in some cases. For example, staying in the current position for safety is extremely conservative. This method [325] shows the near-optimal cumulative reward under some assumptions, whereas it cannot achieve the near-optimal values while guaranteeing safety constraints.

Different from CPO, a first-order policy optimization method [193] is provided based on interior point optimization (IPO) [45], in which the logarithmic barrier functions are leveraged to satisfy safety constraints. The method is easy to implement by tuning the penalty function. Although the empirical results on MuJoCo [6] and grid-world environments [64] have demonstrated their method's effectiveness, the theoretical analysis to guarantee the performance is still needed to be provided.

Unlike the above CMDP optimization, a meta algorithm [221] is proposed to solve safe RL problems with general convex constraints [46]. They can, in particular, solve CMDP problems with a small number of samples in reward-free settings. The algorithms can also be used to solve approachability problems [38, 220]. Although the theoretical results have shown their method's effectiveness, practical algorithms and experiments ought to be proposed and carried out to further evaluate their method. In [240], an attempt is made to solve safe RL using the trajectory probability in a safe set. The probability invariance is positive, and constraint gradients are obtained using the policy parameters. More importantly, the related problem can have an arbitrarily small duality gap. However, this method may also encounter the stability problems of primal-dual methods, and the saddle points using Lagrangian methods may be unstable [237].

Comparable to the above primal-dual settings, Ghosh *et al.* [111] have

[7]The SNO-MDP represents the Safe Near-Optimal MDP.

developed a model of safe RL based on least-squares value iteration-upper confidence bound [148] in primal-dual settings, which is suitable for large-scale optimization, and the regret bound and safety violation are analyzed. While similar methodologies have been proposed, they are predominantly utilized within tabular settings, as exemplified by the work of Wei et al. [333]. Furthermore, Xiong *et al.* [340] introduce a step-wise safe RL method based on upper confidence bound value iteration [21]. This method employs optimistic estimations of transition kernels and cost functions, thereby enhancing its capacity to ensure safety in reward-free scenarios. Notably, Xiong *et al.* also provide analyses of safety violations and regret bounds, facilitating the extension of this method to additional contexts such as multi-robot systems. Although these methods demonstrate improved performance over related baselines, further exploration into their applications across diverse domains, such as safe multi-robot control and planning, could yield even more beneficial insights. In recent years, primal-based methods [343] have emerged as powerful alternatives to primal-dual-based approaches [6, 344], demonstrating remarkable performance in robotics applications. Notable examples include CRPO [343] and PCRPO [119]. A significant advantage of primal-based methods is that they do not require tuning dual parameters and are not influenced by initialization [343]. Specifically, CRPO introduced a primal-based framework for safe RL, but it may have experienced oscillations between reward and safety optimization, potentially undermining its performance. To address this issue, PCRPO introduced a soft policy optimization technique with gradient manipulation, exhibiting superior performance compared to CRPO and state-of-the-art primal-dual-based methods such as PCPO [348], CUP [344], and PPOLag [260].

3.2.2. Control theory-based approaches

In contrast to safe policy optimization such as meta algorithms, CPO and IPO based on model-free RL, control theory is are also leveraged to ensure RL safety. For example, the first Lyapunov functions used for a safe RL may be [248], in which the agent's actions are constrained by applying the control law of Lyapunov functions to learning systems and removing the unsafe actions in the action set. The experiments have demonstrated that their method can achieve safe actions for the control problems in their study. However, the method requires the knowledge of a Lyapunov function in advance. If the environment dynamics model is unknown, it may be difficult to address safe RL problems with this method. Unlike [248], in [62] and [63], Chow *et al.*

propose several safe RL methods based on Lyapunov functions in discrete and continuous CMDP settings, respectively, where Lyapunov functions are used for safe RL to guarantee safe policy and learning stability. The methods can guarantee safety during training, and the Lyapunov functions can be designed by a proposed Linear Programming algorithm. However, the training stability and safety using Lyapunov functions still need to be improved, and more efficient algorithms in the setting may need to be proposed.

Apart from Lyapunov functions based safe RL, CBFs are also developed in model-free settings in recent years [351, 323], [351] develop a model-free safe RL method based CBFs, in which they learn the policy and CBFs with data-driven methods, which can help to reduce reliance on the environment models. Similarly, [323] proposes a safe RL method for power system control with CBFs, in which they integrated CBFs into reward functions to search for safe policy. model free control barrier function, [207], [358]. Moreover, [57] leverages a model-free RL and CBFs to guarantee learning safety while improving learning efficiency, in which CBFs are used to constrain the search space to enhance learning safety and efficiency based on a GP learning model. Moreover, [208] also leverages CBFs to ensure learning safety with learned dynamics knowledge, where they provide convergence and stability analysis, and their method outperforms the baselines in their experiments.

Safety layer methods have been proposed in recent years, which are more directly to ensure RL safety than safe policy optimization methods that based on cost value optimization such as CPO and IPO. We categorize safety-layer-based approaches into control theory methods since they can be considered as a control filter in control theory.

For example, Pham *et al.* propose an OptLayer [250] method, in which they leverage stochastic control policies to attain the reward performance, and a layer of the neural network is integrated to pursue safety during applications. The real-world applications also demonstrate the effectiveness of the safety layer. Qin *et al.* [256] propose the DCRL (Density Constrained Reinforcement Learning) method that is to optimize the reward and cost from the perspective of an optimization criterion, in which they consider the safety constraints via the duality property [73, 228, 229, 306] with regard to the state density functions, rather than the cost functions, reward functions and value functions [6, 74, 82]. This method lies in model-free settings whereas Chen *et al.* [56] provide similar methods in model-based settings. A-CRL (state Augmented Constrained Reinforcement Learning) method [52] is proposed to address a CMDP problem whereby the optimal policy may not be achieved

via regular rewards. Their method focuses on solving the monitor problem in CMDP while the dual gradient descent is used to find the feasible trajectories and guarantee safety. Nonetheless, OptLayer [250], A-CRL [52] and DCRL [256] all lack convergence rate analysis.

3.2.3. Gaussian processes-based approaches

Relative to safe policy optimization methods that attend more to cost values, Lyapunov function-based techniques, which emphasize safe actions, and Safety layer approaches, GP methodologies primarily concentrate on the safe exploration via modeling the safe states. In a GP of model-free settings, Sui *et al.* [297] use a GP method to present the unknown reward function from noise samples, the exploration by leveraging a GP method is improved to reduce uncertainty and ensure agent safety. More particularly, the GP method is used to predict unknown function, and guide the exploration in bandit settings which do not need to state transitions. Their real-world experiments in movie recommendation systems and medical areas indicate that their method can achieve near-optimal values safely. Like Sui *et al.* [297], Turchetta *et al.* [319] also leverage a GP method to approximate unknown functions prior for safe exploration. Nevertheless, they focus more on finite MDP settings considering explicitly reachability. Nonetheless, the method may not optimize reward objectives while considering safety. Wachi *et al.* [326] represent unknown reward and cost functions with GP methods to ensure safety with probability and optimize reward. Furthermore, the safe, unsafe, and uncertain states are denoted for agent optimistic and pessimistic exploration, and their method can adapt the trade-off between exploration and exploitation. Nevertheless, the convergence guarantee with finite-time rates, optimization for multiple and heterogeneous objectives may need to be provided.

Although GP methods have shown impressive performance with regard to RL safety, most of them ensure safety with probability. How to rigorously guarantee RL safety during exploration still remains open.

3.2.4. Formal methods-based approaches

Distinct from employing GP methods to model the safe state, Hasanbeig et al. [129] propose a safety-enhanced RL approach that leverages LTL. In this novel methodology, logical formulas serve as constraints during the exploration phase of policy synthesis. Although this method demonstrates notable safety performance, it is imperative to carefully define the logical

constraints to ensure safe exploration and effectively balance the trade-off between safety performance and reward acquisition. Murugesan *et al.* [227] employ a formal tool, satisfiability modulo theories [66], as a safety verification layer to enhance learning safety. Furthermore, Hasanbeig *et al.* [1] encode properties of LTL into the reward function through reward engineering using a limit deterministic Büchi automaton [285] to ensure safety. While the aforementioned methods exhibit impressive performance in ensuring safety in most tasks, their deployment in real-world applications still needs to be further investigated.

3.3. Safe Multi-Agent Reinforcement Learning

Safe RL has received increasing attention both from academia and industry. However, most current RL methods are based on the single-agent setting. Safe MARL is still a relatively new area that has emerged in recent years. Little research has yet been carried out that considers the safe multi-agent RL, which can be seen as a multi-agent CMDP problem. Safe multi-agent RL not only needs to consider the ego agent's safety and other agents' safety, but also needs to take into account the ego agent's reward and other agents' reward. In this section, we briefly introduce the safe multi-agent RL problem formulation (since merely a few safe MARL methods are developed up to date, MARL problem formulation is not given in Section 2), and some safe multi-agent RL methods are analyzed in detail.

3.3.1. Problem Formulation of Safe Multi-Agent Reinforcement Learning

In this section, We mainly investigate MARL in a fully cooperative setting, and it's can be seen as a multi-agent constrained Markov game considered as the tuple $\langle \mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}, \mathrm{P}, \rho^0, \gamma, R, \boldsymbol{C}, \boldsymbol{c} \rangle$. A set of agents is present as $\mathcal{N} = \{1, \ldots, n\}$, $\mathcal{S}$ is the state space, $\boldsymbol{\mathcal{A}} = \prod_{i=1}^{n} \Pi^i$ is the product of agents' action spaces, known as the joint action space, $P : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to [0, 1]$ is a function of probabilistic state transition, the initial state distribution is ρ^0, $\gamma \in (0, 1)$ is a discount factor, $R : \mathcal{S} \times \boldsymbol{\mathcal{A}} \to \mathbb{R}$ is the joint reward function, $\boldsymbol{C} = \{C_j^i\}_{1 \le j \le m^i}^{i \in \mathcal{N}}$ is the set of cost functions $C_j^i : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}$ of individual agents, and finally the set of cost-constraining values is given by $\boldsymbol{c} = \{c_j^i\}_{1 \le j \le m^i}^{i \in \mathcal{N}}$. At time step t, the agents are in state s_t, and every one of them performs an action according to its policy $\pi^i(\mathrm{a}^i|\mathrm{s}_t)$. Together with other agents' actions, this results in a joint action $\mathbf{a}_t = (\mathrm{a}_t^1, \ldots, \mathrm{a}_t^n)$ drawn from the joint policy $\boldsymbol{\pi}(\mathbf{a}|\mathrm{s}) = \prod_{i=1}^{n} \pi^i(\mathrm{a}^i|\mathrm{s})$. The agents receive the reward $R(\mathrm{s}_t, \mathbf{a}_t)$,

Model-Free Safe RL	Features	Methods	Convergence Analysis
[221, Miryoosefi and Jin (2021)]	Provide a meta algorithm to solve CMDP problems using model free methods.	Reward-free methods for CMDP.	NO.
[240, Liang et al., (2022)]	Provide analysis to primal-dual gap can be arbitrarily small, and can converge for any small step-size η_λ.	Primal-dual methods.	YES, by [88].
[240, Achiam et al., (2017)]	Constrained policy optimisation.	Trust region with safety constraints.	NO.
[240, Liu et al., (2020)]	PPO with logarithmic barrier functions.	Primal-dual methods.	NO.
[62, 63, Chow et al., (2018,2019)]	An effective Linear Programming method to generate Lyapunov functions, and the algorithms can ensure feasibility, and optimality for discrete and continuous system under certain conditions.	Lyapunov function methods.	NO.
[348, Yang et al., (2020)]	Two-step optimisation, first reward, second safety.	Trust region with safety constraints.	YES.
[361, Zhang et al.,(2020)]	First order optimisation with two-step optimisation, first nonparameterized policy, second parameterized policy.	Primal-dual and Trust region methods.	YES.
[327, Wagener et al., (2021).]	Two-step optimisation with advantage. First, MDP optimisation Second, CMDP optimisation by chance optimisation.	An intervention-based method.	NO.
[325, Wachi et al.,(2020)]	Optimse CMDP under unknow safety constraints. First, search safe policy by expanding safe region. Second, optimse reward in safe regions.	Gaussian process optimisation.	YES.
[250, Pham et al., (2018)]	Select safe action by a constrained neural network, and carry out experiments in simulation and real-world environments.	Trust region method with a neural network constraints.	NO.

Table 2: Model-free safe RL analysis.

and each agent i pays the costs $C^i_j(\mathrm{s}_t, \mathrm{a}^i_t)$, $\forall j = 1, \ldots, m^i$. The whole system moves to the state s_{t+1} with probability $\mathrm{P}(\mathrm{s}_{t+1}|\mathrm{s}_t, \mathbf{a}_t)$. The agents' goal is to maximise the joint return, given by

$$J(\boldsymbol{\pi}) = \mathbb{E}_{\mathrm{s}_0 \sim \rho^0, \mathbf{a}_{0:\infty} \sim \boldsymbol{\pi}, \mathrm{s}_{1:\infty} \sim \boldsymbol{P}} \left[\sum_{t=0}^{\infty} \gamma^t R(\mathrm{s}_t, \mathbf{a}_t) \right], \tag{11}$$

while obeying the safety constraints

$$J_j^i(\boldsymbol{\pi}) = \mathbb{E}_{\mathrm{s}_0\sim\rho^0,\mathbf{a}_{0:\infty}\sim\boldsymbol{\pi},\mathrm{s}_{1:\infty}\sim\boldsymbol{P}}\left[\sum_{t=0}^{\infty}\gamma^t C_j^i(\mathrm{s}_t,\mathrm{a}_t^i)\right]. \tag{12}$$

The state-action value function, as well as the state-value function are defined as

$$Q_{\boldsymbol{\pi}}(s,\boldsymbol{a}) = \mathbb{E}_{\mathrm{s}_{1:\infty}\sim\mathrm{P},\mathbf{a}_{0:\infty}\sim\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t R(\mathrm{s}_t,\mathbf{a}_t)\Big|\mathrm{s}_0=s,\mathbf{a}_0=\boldsymbol{a}\right],$$
$$V_{\boldsymbol{\pi}}(s) = \mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}}\left[Q_{\boldsymbol{\pi}}(s,\mathbf{a})\right],$$

respectively.

In multi-agent settings, the contribution of actions of subsets of the agent set via the multi-agent state-action value function and state value function is introduced as Q-function and V-function:

$$\begin{cases} Q^{\pi^{i_{1:h}},\boldsymbol{\pi}^{-\boldsymbol{i_h}}}\left(s_t,a_t^{i_{1:h}},\boldsymbol{a_t^{-i_h}}\right) = \mathbb{E}_{s_{t+1},a_{t+1}^{i_{1:h}},\boldsymbol{a_{t+1}^{-i_h}}}\left[\sum_{l=0}^{\infty}\gamma^l R\left(s_{t+l},a_{t+l}^{i_{1:h}},\boldsymbol{a_{t+l}^{-i_h}}\right)\right] \\ V^{\pi_{i_{1:h}},\boldsymbol{\pi}_{-\boldsymbol{i_h}}}(s_t) = \mathbb{E}_{a_t^{i_{1:h}},\boldsymbol{a_t^{-i_h}},s_{t+1}}\left[\sum_{l=0}^{\infty}\gamma^l R\left(s_{t+l},a_{t+l}^{i_{1:h}},\boldsymbol{a_{t+l}^{-i_h}}\right)\right] \end{cases} \tag{13}$$

The advantage for policy optimisation is given as

$$A^{\pi^{i_{1:h}},\boldsymbol{\pi}^{-\boldsymbol{i_h}}}\left(s_t,a_t^{i_{1:h}},\boldsymbol{a_t^{-i_h}}\right) = Q^{\pi^{i_{1:h}},\boldsymbol{\pi}^{-\boldsymbol{i_h}}}\left(s_t,a_t^{i_{1:h}},\boldsymbol{a_t^{-i_h}}\right) - V^{\pi^{i_{1:h}},\boldsymbol{\pi}^{-\boldsymbol{i_h}}}(s_t) \tag{14}$$

Similar to reward, cost for j constraint of any agent i in a multi-agent system:

$$\begin{cases} Q_j^{\pi_{i_{1:h}},\boldsymbol{\pi}_{-\boldsymbol{i_h}}}\left(s_t,a_t^{i_{1:h}},\boldsymbol{a_t^{-i_h}}\right) = \mathbb{E}_{s^{t+1},a_{i_{1:h}}^{t+1},a_{-i_h}^{t+1}}\left[\sum_{l=0}^{\infty}\gamma^l C_j^i\left(s_{t+l},a_{t+l}^{i_{1:h}},\boldsymbol{a_{t+l}^{-i_h}}\right)\right] \\ V_j^{\pi_{i_{1:h}},\boldsymbol{\pi}_{-\boldsymbol{i_h}}}(s_t) = \mathbb{E}_{a_t^{i_{1:h}},a_t^{-i_h},s_{t+1}}\left[\sum_{l=0}^{\infty}\gamma^l C_j^i\left(s_{t+l},a_{t+l}^{i_{1:h}},\boldsymbol{a_{t+l}^{-i_h}}\right)\right] \end{cases} \tag{15}$$

$$A_j^{\pi^{i_{1:h}},\boldsymbol{\pi}^{-\boldsymbol{i_h}}}\left(s_t,a_t^{i_{1:h}},\boldsymbol{a_t^{-i_h}}\right) = Q_j^{\pi_{i_{1:h}},\boldsymbol{\pi}_{-\boldsymbol{i_h}}}\left(s_t,a_t^{i_{1:h}},\boldsymbol{a_t^{-i_h}}\right) - V_j^{\pi^{i_{1:h}},\boldsymbol{\pi}^{-\boldsymbol{i_h}}}(s_t) \tag{16}$$

The goal of a safe MARL algorithm under the feasible set constraint (18) is expressed as equation (17), the optimal policy (19) is that the policy can maximize the agents' reward and satisfy the constrained conditions,

$$\begin{aligned} &\max\ J\left(\pi^{i_{1:h}},\boldsymbol{\pi}^{-\boldsymbol{i_h}}\right), \\ &\text{s.t. } J_j^i\left(\pi^{i_{1:h}},\boldsymbol{\pi}^{-\boldsymbol{i_h}}\right) \le c_j^i \text{ for all } 1\le j\le m,\ 1\le i\le n, 1\le h\le n. \end{aligned} \tag{17}$$

$$\Omega_C = \left\{ \pi^{i_{1:h}}, \boldsymbol{\pi}^{-\boldsymbol{i_h}} : \forall\ 1 \leq j \leq m, 1 \leq i \leq n, 1 \leq h \leq n, J_j^i \left(\pi^{i_{1:h}}, \boldsymbol{\pi}^{-\boldsymbol{i_h}} \right) \leq c_j^i \right\}, \tag{18}$$

$$(\pi_*^{-i_{1:h}}, \boldsymbol{\pi}_*^{-\boldsymbol{i_h}}) = \underset{\pi^{i_{1:h}}, \boldsymbol{\pi}^{-\boldsymbol{i_h}} \in \Omega_C}{\operatorname{argmax}} \{ J \left(\pi^{i_{1:h}}, \boldsymbol{\pi}^{-\boldsymbol{i_h}} \right) \} \quad \forall i. \tag{19}$$

3.3.2. Methods of Safe Multi-Agent Reinforcement Learning

The Multi-Agent Constrained Policy Optimisation (MACPO) algorithm [118] is the first safe model-free MARL algorithm which is developed based on CPO [6] and Heterogeneous-Agent Trust Region Policy Optimisation (HATRPO) [171]; it can guarantee monotonic improvement in reward, while theoretically satisfying safety constraints during each iteration. The algorithm was tested on several challenging tasks, e.g. safe multi-agent Ant tasks, safe multi-agent Lift tasks. However, MACPO is an expensive computation algorithm since the Fisher Information Matrix is computed to approximate conjugate gradients. In contrast to MACPO, MAPPO-Lagrangian [118] is the model-free and first-order safe MARL algorithm that requires less time for computation on most of challenging tasks, and is easily implemented. In addition, by adaptive updating of Lagrangian multipliers, MAPPO-Lagrangian can avoid the problem of the sensitivity to the initialization of the Lagrange multipliers, that other Lagrangian relaxation methods have. Nonetheless, MAPPO-Lagrangian does not guarantee hard safety constraints. [117] presents a soft-constrained policy optimization approach for real-world robot control, which addresses sim-to-real challenges. Despite these advancements, further investigation is required to establish robust theoretical guarantees.

Safe Decentralized Policy Gradient (Dec-PG) [198] using a decentralized policy descent-ascent method is the most closely-related algorithm to MACPO. The saddle point considering reward and cost is searched by a primal-dual framework in a multi-agent system. However, a consensus network used in this method imposes an extra constraint of parameter sharing among neighboring agents, which could yield suboptimal solutions [171]. In contrast, MACPO does not require the assumption of parameter sharing. Furthermore, multi-agent policy gradient methods can suffer from high variance by parameter sharing [172].

Robust MARL [359] is a model-based MARL algorithm, which takes into account the reward uncertainty and transition uncertainty. In their

experiments, they take into account the reward uncertainty by randomizing a nature policy with Gaussian noise, as a constraint to optimize the agent performance. Nevertheless, the payoff assumption for each agent to get a higher payoff when other agents do not satisfy the equilibrium policies, and the equilibrium assumption for global optimization, may be too strong in real-world applications.

Another recently safe MARL work is CMIX [189], which considers peak and average constraints using QMIX [257]. Even though CMIX can satisfy the multi-agent cost constraint in these experiments, it does not provide theoretical analysis to guarantee safety for a multi-agent system. CMIX is also the parameter-sharing algorithm that suffers from problems similar to those of Safe Dec-PG [198].

Based on formal methods [336], Safe MARL via shielding is proposed[91], in which they study the safe MARL problem by one of the formal methods, Linear Temporal Logic (LTL) [251] that is already adopted in safe single-agent settings [9]. Specifically, an attempt is made to ensure the safety of multi-agent systems by leveraging central shielding or factored shielding methods, which are used to specify the safe actions that agents only can take, and also correct the unsafe actions that agents explored. Furthermore, they evaluate the methods based on Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [197] and Coordinating Q-learning (CQ-learning) methods [75]. Although this method can guarantee safety for multi-agent systems to some extent, it requires a prior LTL safety specification and problem-specific knowledge in advance. In a similar way that [91] ensures safety by optimizing the exploration, [284] developed a safe MARL algorithm by adding a safety layer based on safe DDPG [74] and MADDPG [197] for continuous action space settings. As a result, their method greatly enhanced safety on some MARL tasks. However, the method does not guarantee zero violation in all tasks that they carried out.

3.4. Summary of Safe Reinforcement Learning Methods

In this section, we introduce safe RL algorithms from different perspectives, such as value-based safe RL, policy-based safe RL; safe single-agent RL, safe multi-agent RL; model-based safe RL, model-free safe RL. Even though a number of algorithms display impressive performance in terms of reward scores, there is a long way to go toward real-world applications. In addition, based on our investigation, in MARL settings, [176] is one of the earliest methods that present convergence guarantee for multi-agent systems in 2000.

However, only a few algorithms consider safety constraints, and safe MARL is still a relatively new area that requires a lot more attention.

4. Theory of Safe Reinforcement Learning

This section investigates theoretical techniques for analyzing safe RL. Firstly, we introduce the essential theoretical differences between standard RL and safe RL. Secondly, we investigate theories for primal-dual approaches and constrained policy optimization, two prominent methods in theoretical safe RL. Finally, we delve into safe RL sampling complexity and safety violation analysis, and discuss other theoretical frameworks for safe RL. This section aims to comprehensively investigate safe RL theories by clearly delineating the theoretical foundations and analytical techniques. This exploration demonstrates the critical differences from standard RL and highlights the theoretical underpinnings of various safe RL settings and their associated analyses.

4.1. Difference between RL and Safe RL

The standard RL exists a deterministic stationary policy [255], while CMDP may not exist uniformly optimal stationary policies [12], which is one of the main differences between safe RL and standard RL. We present it as follows, and the blog [301] presents a two-state example to illustrate this issue.

Theorem 1. *If $\Pi_C \neq \varnothing$, then there exists a $\pi_\star \in \Pi$ is the optimal policy of CMDP, where the optimal policy $\pi_\star$ belongs to the following two cases:*

- *$\pi_\star$ is a deterministic stationary policy;*
- *or $\pi_\star$ is a randomized stationary policy. Its randomness only occurs on the unique state $\tilde{s} \in \mathcal{S}$, where the policy $\pi_\star$ randomly plays two actions on $\tilde{s} \in \mathcal{S}$. For the state $s \neq \tilde{s}$, the policy $\pi_\star$ is a deterministic stationary policy.*

Additionally, the next theorem shows that checking a safe policy is time-expensive.

Theorem 2. *[96] Checking feasibility in CMDPs among deterministic policies is NP-hard.*

4.2. Theory for Primal-Dual Approaches

A standard way to solve a CMDP problem is the Lagrangian approach [60] that is also known as primal-dual optimization,

$$(\pi_\star, \boldsymbol{\lambda}_\star) = \arg\min_{\boldsymbol{\lambda}\geq \mathbf{0}} \max_{\pi\in\Pi_S} \{J(\pi_{\boldsymbol{\theta}}) - \boldsymbol{\lambda}^\top(\mathbf{c}(\pi) - \mathbf{b})\}\,. \tag{20}$$

Extensive canonical algorithms are proposed to solve problem (20), e.g., [55, 177, 186, 241, 270, 274, 310, 342, 80].

The work [3] presents a policy-based algorithm to solve the CMDP problem with average cost finite states, which is a stochastic approximation algorithm. Furthermore, the work [3] shows the locally optimal policy of the proposed algorithm. Chen et al. [310] propose the Reward Constrained Policy Optimization (RCPO) that is a multi-timescale algorithm for safe RL, and [310] show the asymptotical convergence of RCPO. The main idea of achieving such an asymptotical convergence is stochastic approximation [42, 265] that has been widely used in RL, e.g., [252, 313, 345]. For the global non-asymptotic convergence guarantees, see [41, 36, 320, 2, 61, 185, 130].

Based on Dankin's Theorem and Convex Analysis [32], [241] provides theoretical support to the primal-dual (i.e., the Lagrange multiplier) method with a zero duality gap, which implies that the primal-dual method can be solved precisely in the dual domain. In [177], they consider a framework for off-line RL under some constraint conditions, present a specific algorithmic instantiation, and show the performance guarantees that preserve safe learning. Moreover, due to the off-line RL, according to the off-policy policy evaluation, the work [177] shows the error bound with respect to Probably Approximately Correct (PAC) style bounds [183]. The work [270] is proposed to merge the theory of constrained CMDP with the theory of robust Markov decision process (RMDP). The RMDP leads to a formulation of a robust constrained Markov decision process (RCMDP), which is the essential formulation to build a robust soft-constrained Lagrange-based algorithm. The work [270] claims the convergence of the proposed algorithm can follow the method [42] because the proposed algorithm can present a time-scale and stochastic approximation. The work [274] firstly turn the cost-based constraints into state-based constraints, then propose a policy improvement algorithm with a safety guarantee. In [274], the authors propose backward value functions that play a role in estimating the expected cost according to the data by the agent, their main idea is policy iteration [272].

4.3. Theory for Constrained Policy Optimization

The work CPO [6] suggests to use a surrogate cost function which evaluates the constraint $J^c(\pi_{\boldsymbol{\theta}})$ according to the samples collected from the current policy $\pi_{\boldsymbol{\theta}_k}$. Although using a surrogate function to replace the cumulative constraint cost has appeared in [74, 90, 103], CPO [6] firstly show their algorithm guarantees for near-constraint satisfaction.

$$\pi_{\boldsymbol{\theta}_{k+1}} = \arg\max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \quad \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A_{\pi_{\boldsymbol{\theta}_k}}(s,a) \right] \tag{21}$$

$$\text{s.t.} \quad J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)} \left[A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a) \right] \leq b, \tag{22}$$

$$\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k}) = \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} [\mathrm{KL}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})[s]] \leq \delta. \tag{23}$$

Existing recent works (e.g., [6, 35, 37, 127, 324, 348]) try to find some convex approximations to replace the terms $A_{\pi_{\boldsymbol{\theta}_k}}(s,a)$ and $\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})$. Concretely, [6] suggest using the first-order Taylor expansion to replace (21)-(22), the second-oder approximation to replace (23). Such first-order and second-order approximations turn a non-convex problem (21)-(23) into a convex problem. This would appear to be a simple solution, but this approach results in many error sources and troubles in practice. Firstly, it still lacks a theory analysis to show the difference between the non-convex problem (21)-(23) and its convex approximation. Policy optimization is a typical non-convex problem [347]; its convex approximation may introduce some errors for its original issue. Secondly, CPO updates parameters according to conjugate gradient [298], and its solution involves the inverse Fisher information matrix, which requires expensive computation for each update. Later, the work [348] proposes PCPO that also uses second-order approximation, resulting in an expensive computation.

Conservative Safety Critics (CSC) [35] provides a new estimator for the safe exploration of RL. CSC uses a conservative safety critic to estimate the environmental state functions. The likelihood of catastrophic failures has been bounded for each round of the training. The work [35] also shows the trade-off between policy improvement and safety constraints for its method. During the training process, CSC keeps the safety constraints with a high probability, which is no worse asymptotically than standard RL.

First Order Constrained Optimization in Policy Space (FOCOPS) [361] and onservative update policy (CUP) methods [344] suggest that the non-convex implementation solving constrained policy optimization contains a

two-step approach: policy improvement and projection. The work [361] has shown the upper boundedness of the worst case for safety learning, where it trains the model via first-order optimization. CUP [344] provides a theoretical analysis extending the bound concerning the generalized advantage estimator (GAE) [280]. GAE significantly reduces variance while achieving a tolerable level of bias, which is one of the critical steps when designing efficient algorithms [346]. In addition, the asymptotic results that safe RL methods can achieve are summarised in Table 3.

Reference	Main Technique	Method	Implementation
[3]	SA [42] and Iterate Averaging [174]	Primal-Dual	Policy Gradient
RCPO [310]	SA and ODE [42]	Primal-Dual	Actor-Critic
[41]	SA [42] and Envelope Theorem [216]	Primal-Dual	Actor-Critic
[241]	Dankin's Theorem and Convex Analysis [32]	Primal-Dual	Actor-Critic
[177]	FQI [226] and PAC [183]	Primal-Dual	Value-based
[270]	SA [42]	Primal-Dual	Policy Gradient
[274]	Policy Iteration [272]	Primal-Dual	Policy Gradient
CPO [6]	Convex Analysis [32]	CPO-based	Policy Gradient
PCPO [348]	Convex Analysis [32]	CPO-based	Policy Gradient
FOCOPS [361]	Non-Convex Analysis	CPO-based	Actor-Critic
CUP [344]	Non-Convex Analysis	CPO-based	Actor-Critic
CSC [35]	On-line Learning [7]	CPO-based	Actor-Critic
SAILR [327]	Back-up Policy	/	Actor-Critic
SNO-MDP [325]	Early Stopping of Exploration of Safety	GP-based[258]	Actor-Critic
A-CRL [52]	On-line Learning	Primal-Dual	Value-based
DCRL [256]	On-line Learning [7]	/	Actor-Critic
Saddle-FD [363]	Primal-Dual	Primal-Dual	Actor-Critic
ReLOAD [225]	Lagrangian	Primal-Dual	Actor-Critic
OPD [76]	/	Primal-Dual	Actor-Critic
Cutting-Plane [113]	Vaidya's Dual Optimizer	Primal-Dual	Actor-Critic
RPGPD [80]	/	Primal-Dual	Actor-Critic

Table 3: Asymptotic results of safe RL (SA denotes Stochastic Approximation, ODE denotes Order Different Equation).

4.4. Sampling Complexity and Safety Violation Analysis

In this section, we review the sampling complexity and safety violation analysis of model-based and model-free safe RL $\mathcal{O}(\epsilon)$-optimality (sample complexity and safety violation of brief introduction that we investigate studies is given in Table 4 and Table 5), where we define a policy π of $\mathcal{O}(\epsilon)$-optimality as follows,

$$J(\pi) - J(\pi_\star) \leq \epsilon, \ \{J^c(\pi) - b\}_+ \leq \epsilon. \tag{24}$$

In this section, we review the sampling complexity of the algorithms that match $\mathcal{O}(\epsilon)$-optimality.

Assumption 1 (Feasibility)**.** *The unknown CMDP is feasible, i.e., there exists an unknown policy π which satisfies the constraints. Thus, an optimal policy exists as well.*

Assumption 2 (Slater Condition,[293])**.** *There exists a vector $\xi \prec 0$, and a policy $\bar{\pi}$ such that*

$$J^c(\bar{\pi}) - b \preceq \xi. \tag{25}$$

Assumption 3 (Linear CMDP)**.** *The CMDP is a linear with a kernel feature map $\psi : \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, for each h, there exists a vector θ_h such that $\mathbb{P}(s'|s,a) = \langle \psi(s,a,s'), \theta_h \rangle$; there exists a feature map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$ and vectors $\theta_{r,h}$ and $\theta_{c,h}$, such that, $r_h(s,a) = \langle \varphi(s,a), \theta_{r,h} \rangle$ and $c_h(s,a) = \langle \varphi(s,a), \theta_{c,h} \rangle$.*

It is worth referring to [20, 175] since this bound helps us to understand the complexity of safe RL algorithms, where the works [20, 175] a lower bound of samples to match $o_{(\epsilon)}$ -optimality as $o\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$, which is helpful for us to understand the RL safety guarantees.

4.4.1. Model-Based Safe Reinforcement Learning

Linear programming and Lagrangian approximation are widely used in model-based safe RL if the estimated transition model is either given or estimated accurately enough [11]. OptDual-CMDP [89] achieves sublinear regret with respect to the main utility while having a sublinear regret on the constraint violations, i.e., the OptDual-CMDP needs $o\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$ to achieve a $o_{(\epsilon)}$-optimality. the Upper-Confidence Constrained Fixed-Horizon RL method (UC-CFH) [151] provides a proximal optimal policy under the probably

Model-Based Learning	Algorithm / Reference	Settings	Iteration Complexity	Safety Violation
Lower bound	[175] [20]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	/
Value-Based	OptDual-CMDP [89]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$
Value-Based	OptPrimalDual-CMDP [89]	Assumption1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$
Value-Based	ConRL [46]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$
Value-Based	OptPess-PrimalDual [191]	Assumption 2	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$
Value-Based	UC-CFH [151]	Assumption 2	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$
Value-Based	OPDOP [81, Theorem 1]	Assumption 2,3	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$
Policy-Based	NPG-PD [82, Theorem 1]	Assumption 2	$\mathcal{O}\left(\frac{1}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{(1-\gamma)^4\epsilon^2}\right)$

Table 4: This table summarizes the model-based state-of-the-art algorithms for safe RL or CMDP.

approximately correctness (PAC) analysis. The main idea of UC-CFH is to consider linear programming method to online learning to design an algorithm to finite-horizon CMDP. Concretely, UC-CFH [151] needs $o\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$ samples, and we should notice that according to [23], Theorem 1 in [151] involves a constant $\boldsymbol{C}$ that is bounded by $\lvert\boldsymbol{\mathcal{S}}\rvert$. OptPess-PrimalDual [191] provides a way to keeps the performance with $o\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$ sampling complexity with a known strictly safe policy. OptPess-PrimalDual [191] also claims that OptPess-PrimalDual shares a higher probability to achieve a zero constraint violation.

An Optimistic Primal-Dual proximal policy OPtimization (OPDOP) method [81] shows a bound concerning the feature mapping and the capacity of the state-action space, which leads to a sampling complexity of $o\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$. Besides, the work [81] claims even if the dimension of state space goes to infinity, the bound also holds, which implies the merit of OPDOP. Similar techniques also be considered by [340]. An upper confidence bound value iteration (UCBVI)-γ method [131] achieves the sampling complexity of $o\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$ that matches the minimax lower bound up to logarithmic factors. The work [82] applies a natural policy gradient method to solve constrained Markov decision processes. The NPG-PD algorithm applies the gradient descent method to learn the primal variable, while learning the primal variable via natural policy gradient (NPG). The work [82] shows the sampling complexity of NPG-PD achieves $o\left(\frac{1}{(1-\gamma)^4\epsilon^2}\right)$. We notice that Theorem 1 of [82] shows a convergence rate independent on $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{A}}$. The work [111] presents

the efficient algorithms for safe RL with linear function approximation.

ConRL [46] obtains a sampling complexity of $o\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^6\epsilon^2}\right)$. To achieves this result, the work ConRL [46] provides an analysis under two settings of strong theoretical guarantees. Firstly, [46] assumes that ConRL has a learning maximization process with the concave reward function, and this maximization falls into a convex expected value of constraints. The second setting is that during the learning maximization process, the resources never exceed specified levels. Although ConRL plays two additional settings, the complexity is still higher than previous methods, at least with a factor $\frac{1}{(1-\gamma)^2}$.

Model-Free Learning	Algorithm / Reference	Settings	Iteration Complexity	Safety Violation
Policy-Based	ConRL [46, Remark 3.5]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^5\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^5\epsilon^2}\right)$
Value-Based	CSPDA [23][8]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$
Value-Based	Triple-Q [333]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^{2.5}\lvert\mathcal{A}\rvert^{2.5}}{(1-\gamma)^{18.5}\epsilon^5}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^{2.5}\lvert\mathcal{A}\rvert^{2.5}}{(1-\gamma)^{18.5}\epsilon^5}\right)$
Value-Based	Reward-Free CRL [221]	Assumption 3	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$
Policy-Based	CRPO [343, Theorem 1]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^7\epsilon^4}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^7\epsilon^4}\right)$
Policy-Based	On-Line NPG-PD [357, Theorem 1]	Assumption 2	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^6\lvert\mathcal{A}\rvert^6}{(1-\gamma)^{12}\epsilon^6}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^6\lvert\mathcal{A}\rvert^6}{(1-\gamma)^{12}\epsilon^6}\right)$
Policy-Based	NPG-PD [82, Theorem 4]	Assumption 2	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{(1-\gamma)^4\epsilon^2}\right)$

Table 5: This table summarizes the model-free state-of-the-art algorithms for safe RL or CMDP.

4.4.2. Model-Free Safe Reinforcement Learning

Model-free safe RL algorithms, including IPO [193], Lyapunov-Based Safe RL [62, 63], PCPO [348], SAILR [327], SPRL [294], SNO-MDP [325], FOCOPS [361], A-CRL [52], CUP [344] and DCRL [256] all lack convergence rate analysis.

The work [82] shows NPG-PD obtains a sublinear convergence rate for both learning the reward optimality and safety constraints. NPG-PD solves the CMDP with softmax policy, where the reward objective is a non-concave and cost objective is non-convex, NPG-PD [82] shows that with a proper design, policy gradient can also obtain an algorithm that converges at a sublinear rate. Concretely, the Theorem 4 of [82] shows NPG-PD achieves

[8]The generative model is used, which needs additive samples to create a generative model.

the sampling complexity of $o\left(\frac{|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^4\epsilon^2}\right)$, and we should notice that in Theorem 4 of [82], $|\mathcal{S}|^2|\mathcal{A}|^2$ samples are necessary for the two outer loops. Later, the work [357] extends the critical idea of NPG-PD and proposes an online version of NPG-PD that needs the sample complexity of $o\left(\frac{|\mathcal{S}|^6|\mathcal{A}|^6}{(1-\gamma)^{12}\epsilon^6}\right)$, where we show the iteration complexity after some simple algebra according to [357, Lemma 8-9]. Clearly, online learning NPG-PD [357] needs additional $\mathcal{O}(\epsilon^{-4})$ trajectories than NPG-PD [357].

The work [343] proposed a primal-type algorithmic framework to solve SRL problems, and they show the proposed algorithm needs $o\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^7\epsilon^4}\right)$ sample complexity to obtain $\mathcal{O}(\epsilon)$-optimality, where we notice that the inner loop with $\kappa_{\text{in}} = o\left(\frac{T}{(1-\gamma)|\mathcal{S}||\mathcal{A}|}\right)$ iteration is needed [343, Theorem 3].

The work [23] proposes the CSPDA algorithm needs the sample complexity of $o\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\epsilon^2}\right)$. However, the inner loop of their Algorithm 1 needs an additional generative model. Triple-Q [333] needs the sample complexity of $o\left(\frac{|\mathcal{S}|^{2.5}|\mathcal{A}|^{2.5}}{(1-\gamma)^{18.5}\epsilon^5}\right)$. We show this iteration complexity according to a recent work [23]. Since the work [333] study the finite-horizon CMDP, we believe their Triple-Q plays at least $o\left(\frac{|\mathcal{S}|^2|\mathcal{A}|^2}{\epsilon^5}\right)$, which is still higher than NPG-PD [82] at least with a factor $o(\epsilon^{-3})$. The work [221] proposes a safe RL algorithm that needs $o\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\epsilon^2}\right)$. It is noteworthy that we show the sample complexity here for the worst-case of constraint violation shown in [221] reaches $o\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^4\epsilon^2}\right)$, if the number of constraint function is greater than $|\mathcal{S}|$.

4.5. Other Theory Techniques

SAILR [327] shows a theory of safety guarantees for both development and training, where SAILR does not keep the intervention mechanism after the process of learning. Besides, SAILR [327] also shows the comparison between the capacity of reward learning and optimal safety constraints. SNO-MDP [325] shows how to explore the CMDP, where the safety constraint is unknown to the agent, and provides a theoretical analysis of the policy improvement and the safety constraint under some regularity assumptions A-CRL [52] tries to solve general problems in safe RL, where it augments the state by some Lagrange multipliers, and it reinterprets the Lagrange multiplier method as the dynamics portion. The work [256] shows that the DCRL learns a near-optimal policy while keeping a bounded error even if it meets the imperfect process of learning. All of those methods try to analyze learning from different safety settings or typical tasks.

5. Applications of Safe Reinforcement Learning

RL applications for challenging tasks have a long tradition. Some RL methods are used to solve complex problems before neural network learning arises. For example, TD learning is used to solve backgammon playing problems [307, 308], job-shop scheduling problems, elevator group control problems [71], and a stochastic approximation algorithm with RL properties is utilized to solve pricing options for high-dimensional financial derivatives in two-player and zero-sum games [318]. However, most of the above methods are on a small scale or have linear settings, and most of the problems they solve are discrete. The policy values are almost approximated to address more challenging tasks for large-scale, continuous, and high-dimensional problems, e.g., using neural networks is currently a widely adopted method to learn sophisticated policy strategies in modern RL. In this section, to investigate the **"Safety Applications"** problem, we introduce safe RL applications, including tabular-setting RL, and modern RL applications, such as autonomous driving, robotics, and recommendation systems.

The application methods share high-level techniques with Section III: methods of safe RL. However, application methods are more focused on specific applications and how the methods are deployed in these domains, such as autonomous driving and robotics. For instance, in autonomous driving applications, greater emphasis is placed on the application environment and how to ensure planning safety for autonomous vehicles with safe RL. In contrast, the safe RL method section primarily discusses technical solutions to ensure learning safety, and these methods can be deployed in various applications, e.g., autonomous driving and robotics.

The distinction between the two sections lies in their scope and emphasis. While Section III explores the theoretical and algorithmic underpinnings of safe RL methods, the application methods section delves into the practical considerations and challenges associated with implementing these techniques in real-world scenarios. This section aims to bridge the gap between theoretical frameworks and their practical implementation, investigating the nuances and domain-specific requirements of different application domains. By introducing the deployment of safe RL methods in specific applications, this section provides valuable insights into the real-world constraints, environmental factors, and safety-critical considerations that must be addressed.

5.1. Safe Reinforcement Learning for Autonomous Driving

More recently, many methods have been proposed for autonomous driving based on modern, advanced techniques. The work [254], proposed by Pomerleau, may be one of the first learning-based methods for autonomous driving, developed in 1989. Gu *et al.* [116] provide a motion planning method for automated driving based on constrained RL. They combine traditional motion planning and RL methods to perform better than pure RL or traditional methods. Specifically, the topological path search [120, 364] and trajectory lane model, which is derived from trajectory units [121, 122, 365], are leveraged to constrain the RL search space. Their method can be used very well for corridor scenarios that consider environmental uncertainty.

In contrast to Gu *et al.* [116], Wen *et al.* [334] provide a parallel safe RL method for vehicle lane-keeping and multi-vehicle decision-making tasks by using pure constrained RL methods. They extend an actor-critic framework to a three-layer-neural-network framework by adding a risk layer for autonomous driving safety. The synchronized strategy is used to optimize parallel policies for better searching viable states and speeding up convergence.

Krasowski *et al.* [169] develop a safe RL framework for autonomous driving motion planning, in which they focus more on the high-level decision-making problems for lane changes of vehicles on highways. Based on the work [169], Wang [328] presents a low-level decision-making method via a safety layer of CBFs [13, 230], and a legal safe control method by following traffic rules to ensure motion planning safety for autonomous driving in highway scenarios. Different from Wang's method [328] using CBFs, Cao *et al.* [53] improve the safety of autonomous driving in low-level decision-making settings by integrating a rule-based policy, e.g., a Gipps car-following model [112], into RL framework, and a Monte Carlo tree searching method [48] is used to generate their RL framework policies. Although safe RL for low-level decision-making has been very successful, it is still unable to guarantee autonomous driving safety in complex environments, especially for multiple dynamic and uncertain obstacles.

Mirchevsk *et al.* [218] leverage a Q learning method [331] and a tree-based ensemble method [110] used as a function approximator, to achieve high-level control for lane changes in highway scenarios. Their method has shown impressive performance by reducing collision probability. Nevertheless, this method may only be suitable for two-lane changing environments, since one-lane change options are only considered in the environments at any time. Furthermore, Mirchevsk *et al.* [219] use formal methods [243] to guarantee

safety when they use RL for the safe and high-level planning of autonomous driving in autonomous lane changes. Therefore, their method can be used for more complex environments compared to the work [218], and their method displays good performance in highway scenarios with an arbitrary number of lanes. They also integrate safety verification into RL methods to guarantee agent action safety.

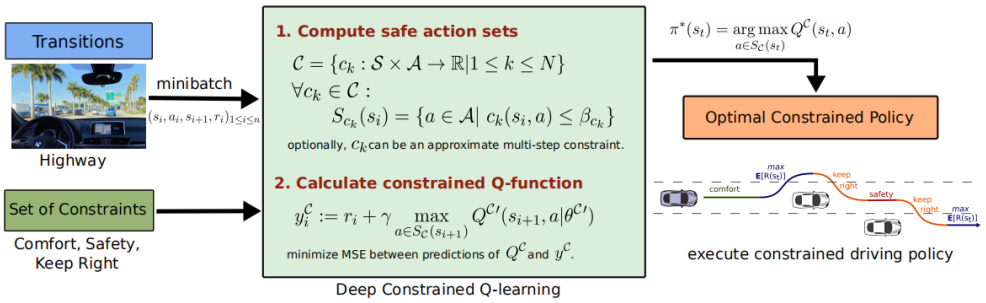


Figure 2: The framework of constrained deep Q-learning for autonomous driving (The figure is adapted with permission from [154]).

In [155], similar to Mirchevsk *et al.* [219], they introduce a verification method for RL safety, more particularly, they verify the action safety. The policy can be learned adaptively in a distributional RL framework. Isele *et al.* [145] use prediction methods to render safe RL exploration for intersection behaviors during autonomous driving. Remarkably, they can constrain agents' actions by prediction methods in multi-agent scenarios, where they assume other agents are not adversarial and an agent's actions are generated by a distribution. Kendall *et al.* [157] provides a model-free based RL method, which is combined by variational autoencoder [160, 262] and DDPG [187]. Their method may be one of the first to implement real-world vehicle experiments using RL, in which they use logical rules to achieve autonomous driving safety, and use mapping and direct supervision to navigate the vehicles.

Kalweit *et al.* [154] develop an off-policy and constrained Q learning method for high-level autonomous driving in simulation environments. They use the transportation software, SUMO [196], as a simulation platform and the real HighD data set [168] to verify the effectiveness of their methods. Specifically, they constrain the agent's action space when the agent performs a Q value update; the safe policy is then searched for autonomous driving (see Figure 2). Different from the above perspectives, Atakishiyev *et al.* introduce some Explainable Artificial Intelligence (XAI) methods, and a framework

for safe autonomous driving [18, 19], in which Explainable Reinforcement Learning (XRL) for choosing vehicle actions is mentioned. Although XRL can be helpful in promoting the development of safe and trustworthy autonomous systems, this topic has just been studied with regard to safe RL, and the relevant research is not remarkably mature.

5.2. Safe Reinforcement Learning for Robotics

Some learning methods for robot applications have shown excellent results [161, 214, 337]. However, the methods do not explicitly consider the agent's safety as an optimization objective. There are a number of works that apply RL methods to simulation robots or real-world robots. However, most of them do not take safety into account during the learning process. For the purpose of better applications using RL methods, we need to figure out how to design safe RL algorithms to achieve better performance for real-world applications. Safe RL is a bridge that tries to improve the safe learning efficiency and connects the RL simulation experiments to real-world applications in robotics.

In [292], Slack *et al.* use an offline primitive learning method, called SAFER, to improve safe RL data efficiency and safety rate. However, SAFER has not theoretical safety guarantees. They collected safe trajectories as a safe set by a scripted policy [291], and applied the safe trajectories to a learning process. In terms of safety and success rate, their method has achieved better performance in PyBullet [69] simulation experiments than other baselines, which are demonstration methods for safe RL (see Figure 3).

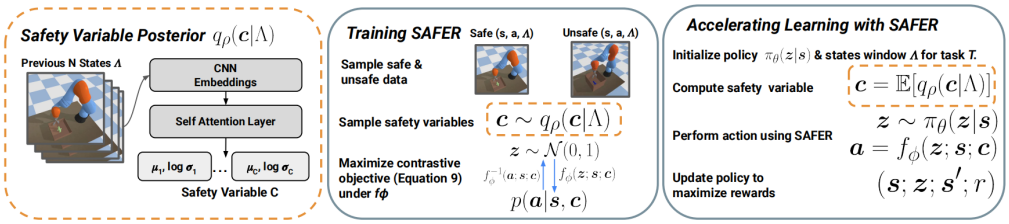


Figure 3: The framework of SAFER (The figure is adapted with permission from [292]).

To facilitate the deployment of safe RL in real-world scenarios, beyond merely simulation-based experiments, the fully differentiable OptLayer [250] is developed to ensure safe actions that the robots can only take. More importantly, they implement their methods in real-world robots using a 6-DoF

industrial manipulator and have received significant attention. However, the method may be limited in high-dimensional space for robot manipulations since the OptLayer may not be able to optimize policies efficiently in complex tasks, especially in high-dimensional space. Furthermore, Garcia and Fernandez [106] present a safe RL algorithm based on safe exploration, in which they develop a smoother and continuous risk function for safe exploration. It can guarantee a monotonic increase, and the risk function is used to help follow a baseline policy. Building on this framework, subsequent work [108] has successfully adapted this risk function-based algorithm for practical applications in real-world robotics.

In the binary step-risk function, a case base $B = \{c_1 \ldots, c_\eta\}$ composed of cases $c_i = (s_i, a_i, V(s_i))$, state risk s is defined by the following equation (26), where $\varrho^{\pi_B^\theta}(s) = 1$ holds if $s \in \Upsilon$; if the state s does not have any relationship with any cases, the state is unknown, it is as $s \in \Omega$, then $\varrho^{\pi_B^\theta}(s) = 0$.

$$\varrho^{\pi_B^\theta}(s) = \begin{cases} 0 & \text{if } \min_{1 \le j \le \eta} d(s, s_j) < \theta \\ 1 & \text{otherwise} \end{cases} \tag{26}$$

The continuous risk function in work [108] is as following: with a state s, a case base $B = \{c_1, \ldots, c_\eta\}$ composed of cases $c_i = (s_i, a_i, V(s_i))$, the risk for each state s is defined by the following function (27). The function can help to achieve a smooth and continuous transition between safe and risky states in their paper.

$$\rho^B(s) = 1 - \frac{1}{1 + e^{\frac{k}{\theta}\left(\left(\min_{1 \le j \le \eta} d(s, s_j) - \frac{\theta}{k}\right) - \theta\right)}} \tag{27}$$

Apart from the risk function (27), the work [108] also implements its algorithm for a real-world robot, a NAO robot [114], as shown in Figure 4.

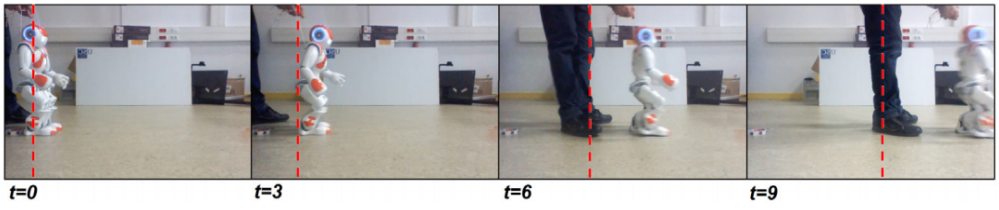


Figure 4: The walking NAO robot (The figure is adapted with permission from [108]).

In contrast to safety layer methods or policy optimization, a series of safe RL methods grounded in control theory have been proposed for robot learning. Notably, Perkins and Barto [248] employ Lyapunov functions, which are developed over a century ago [153], traditionally used to ensure the stability of controllers [282]. They specifically utilize a Lyapunov function to constrain the action space, effectively ensuring the safety of all policies and maintaining agent performance. Moreover, they formulate a set of control laws with predefined Lyapunov domain knowledge. Their approach, applying a Lyapunov-based safe RL method to pendulum tasks using a robot arm in simulated environments, may be among the initial applications of Lyapunov functions for safe RL in robotics. Although Lyapunov functions can enhance system safety and stability, a significant challenge remains: designing a function that fulfills all policy safety requirements necessitates a deep understanding of the system dynamics and the specific domain knowledge of Lyapunov functions.

Similarly, several safe RL methods that require system models are developed in recent years. In particular, Thomas *et al.* [311] leverage a model-based RL to achieve the agents' safety by incorporating the model knowledge. Specifically, they use the agents' dynamics to anticipate the trajectories of the next few steps, and thus prevent agents from entering unsafe states or performing unsafe actions. Based on their method, they apply the proposed method for MuJoCo robot control in simulation environments. Their method may be more suitable for short-horizon trajectories. However, if it encounters a large-scale horizon, the method may not work well since it needs to plan the next few steps quickly. Furthermore, in [190], Liu *et al.* provide a safe exploration method for robot learning on the constrained manifold. More specifically, the robot models and manifold constraints in tangent space are utilized to help ensure robot safety during the RL exploration process. Their method can leverage any model-free RL methods for robot learning on the constrained manifold, since the constrained problem is converted as an unconstrained problem in tangent space, and their method can search for policy on the exploration of safe regions. Nonetheless, an accurate robot model and tracking controller are required in their method, which may not be suitable for real-world applications.

5.3. Safe Reinforcement Learning for Other Applications

Apart from autonomous driving and robotics of safe RL applications, safe RL is also adopted to ensure safety in recommender systems [290], video

compression [205], video transmissions [338], wireless security [199], satellite docking [87], edge computing [339], chemical processes [276] and vehicle schedule [25, 184], and so on. In recommender systems, for example, Singh *et al.* [290] deploy safe RL to optimize the healthy recommendation sequences of recommender systems by utilizing a policy gradient method algorithm on the Conditional Value at Risk (CVaR) method [304], whereby they optimize positive feedback while constraining cumulative exposure of health risk. In video compression, Mandhane *et al.* [205] leverage the Muzero [278], one of the alpha series algorithms, to solve the safe RL problem in video compression. More specifically, as shown in function (28), they optimize the encoded video quality by maximizing quantization parameters (QP) via policy learning, while satisfying the Bitrate constraints. Their experiments have proven that their method can achieve better performance than traditional methods and related modern machine learning methods on the YouTube UGC dataset [330]. However, the method may not be easily scalable for large-scale datasets.

$$\max_{\text{QPs}} \text{Encoded Video Quality} \;\; \text{s.t.} \;\; \text{Bitrate} \;\leq\; \text{Target} \tag{28}$$

In wireless security [199], based on the Inter-agent transfer learning method [72], Lu *et al.* develop a safe RL method for wireless security using a hierarchical structure. More specifically, the target Q network and E-networks with CNN [115] are used to optimize the stability of policy exploration, and reduce the risk of the policy exploration, ultimately enhancing the wireless security in UAV communication against jamming.

5.4. Summary of Applications

In this section, we analyze safe RL methods for autonomous driving and robotics, whereby guaranteeing safety and improving reward simultaneously when the agent is learning is a challenging problem. Some methods are proposed to deal with this problem, such as model-based safe RL to plan safe actions [311], Lyapunov function to guarantee the safety of agents, predefined baseline policy for safe exploration [106], formal verification for safe autonomous driving [219], constrained Q learning for high-level vehicle lane changes [154], etc. Although these methods have been very successful, one major problem that remains is how to rigorously guarantee safety during exploration and retain the reward of monotonic improvement, and how to guarantee stability and convergence when safe RL methods are applied to real-world applications. In addition, we investigate the different types of robot applications using different methods in Table 6.

Methods	Experimental Types	Robots
A genetic algorithm [316]	Simulation experiments	Webots mobile robots [332]
A Particle Swarm Optimization algorithm [232]	Simulation experiments	NAO robots [201]
A learning algorithm [214]	Real-world experiments	The Aldebaran Nao of Humanoid robots [354]
An iterative optimization algorithm [95]	Real-world experiments	The Aldebaran Nao of Humanoid robots [354]
A kinetics teaching method [235]	Real-world experiments	NAO robots [201]
A policy gradient method [163]	Simulation experiments	The Webots simulation package [215]
A Deep RL method [85]	Simulation Experiments	MuJoCo robots [315]
A Neuroevolutionary RL method [165]	Simulation Experiments	Helicopter control [22]
A policy search method [161]	Real-world experiments	A Barrett robot arm

Table 6: Different types of robot applications using different methods.

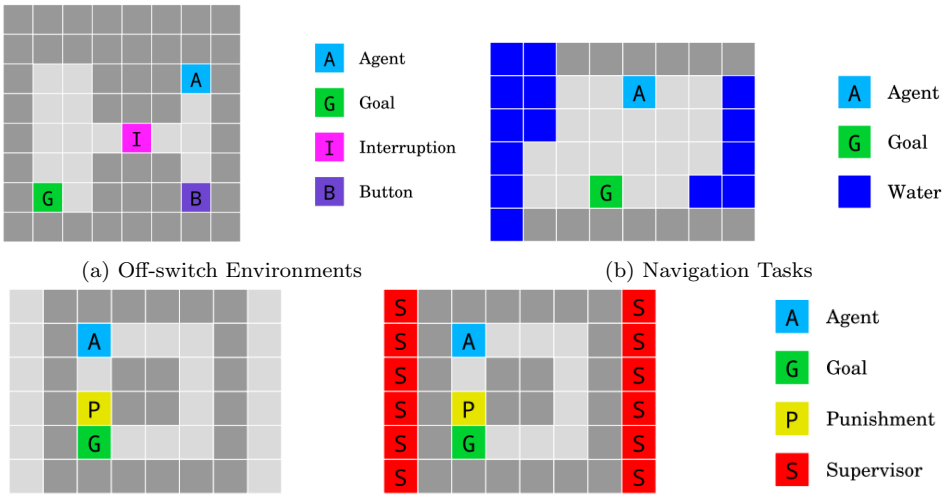
6. Benchmarks of Safe Reinforcement Learning

Several safety benchmarks for safe RL have been developed, and various baselines have been compared on the safe RL benchmarks [259]. More importantly, the Safe RL benchmarks, including single-agent and multi-agent benchmarks, have made massive contributions to the RL community and helped safe RL move toward real-world applications. In this section, we investigate the popular safe RL benchmarks and try to answer the **"Safety Benchmarks"** problem.

6.1. Benchmarks of Safe Single-Agent Reinforcement Learning

6.1.1. AI Safety Gridworlds

AI Safety Gridworlds [181] [9] is a kind of 2-D environment that is used to evaluate safe RL algorithms. All of the environments are based on the 10X10 grids. An agent is arranged in one cell of the grid, and obstacles are arranged in some cells. The action space is discrete in AI safety Gridworlds. An agent can take action from action space $A = \{right, left, up, down\}$, as shown in Figure 5.



(a) Off-switch Environments

(b) Navigation Tasks

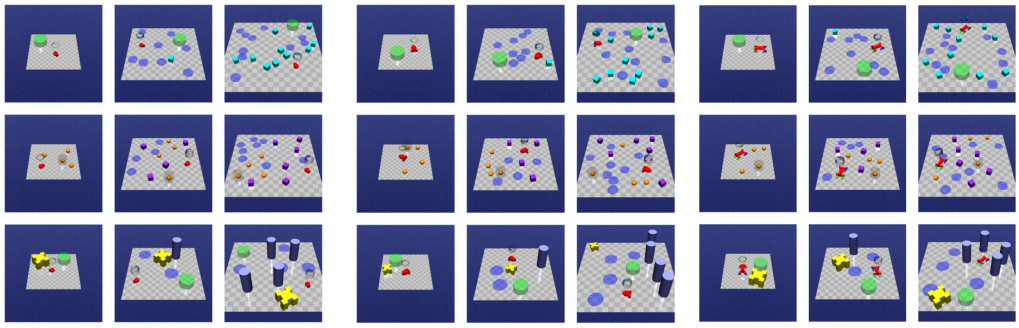
(c) Absent Supervisor Tasks

Figure 5: Tasks of AI safety gridworlds. (a) Off-switch tasks. An agent needs to reach the goal position and has to pass through cell I; when the agent passes cell I, it will be stopped with 50% probability, if the agent goes to cell B, it can switch off the interruption cell I and go to goal cell G with no interruption. (b) Navigation tasks in a island. The agent should not touch water cells, or it will be punished and emit costs. (c) Absent supervisor tasks. In these tasks, the agent needs to reach cell G by a shorter distance. The supervisor will appear with 50% probability if the agent passes cell P. If supervisors S are present, it will be punished and emitted costs, or there will be no punishment (Adapt the figures with permission from [181]).

[9] https://github.com/deepmind/ai-safety-gridworlds.git

6.1.2. Safety Gym

Safety Gym [259] [10] is based on Open AI Gym [47] and MuJoCo [315] environments. It also takes into account 2-D environments in different tasks (as shown in Figure 6), e.g., a 2-D robot such as a Point robot or a Car robot or a Doggo robot can turn and move to navigate a goal position while avoiding crashing into unsafe areas in a 2-D plane. Moreover, the robot's actions are continuous. There are many kinds of costs in the Safety Gym. For example, the robot has to avoid crashing into dangerous areas, non-goal objects, immobile obstacles, and moving objects. Otherwise, costs will be incurred.



(a) Point Envs (b) Car Envs (c) Doggo Envs

Figure 6: Images of Safety Gym environments (Adapt the figure with permission from [259]).

6.1.3. Safe Control Gym

Aiming at safe control and learning problems, Yuan *et al.* propose safe control gym [353] [11], an extension benchmark of OpenAI Gym [47], which integrates traditional control methods [50], learning based-control methods [134] and reinforcement learning methods [125, 281] into a framework, in which model-based and data-based control approaches are both supported. They mainly consider the cart-pole task, 1D, and 2D quadrotor tasks, tasks of —stabilization and trajectory tracking in their environments. Compared with Safety Gym [259] and AI safety gridworlds [181], safe control gym [353] may be more suitable for sim-to-real research, since they offer numerous options

[10] https://github.com/openai/safety-gym.git
[11] https://github.com/utiasDSL/safe-control-gym.git

for implementing non-idealities that resemble real-world robotics, such as randomization, dynamics disturbances and also support a symbolic framework to present systems' dynamics and constraints (see Figure 7).

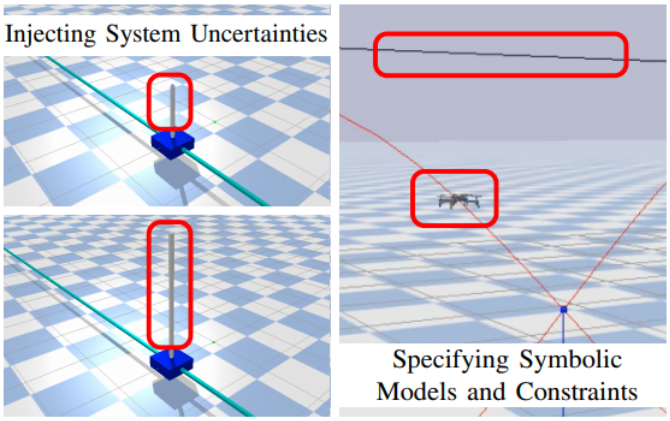


Figure 7: Images of safe control gym with symbolic dynamics constraints in keep tracking and stabiliztion tasks (The figure is adapted with permission from [353]).

6.2. Benchmarks of Safe Multi-Agent Reinforcement Learning

In recent years, three safe multi-agent benchmarks have been developed by [118], namely Safe Multi-Agent MuJoCo (Safe MAMuJoCo), Safe Multi-Agent Robosuite (Safe MARobosuite), Safe Multi-Agent Isaac Gym (Safe MAIG), respectively. The safe multi-agent benchmarks can help promote the research of safe MARL.

6.2.1. Safe Multi-Agent MuJoCo

Safe MAMuJoCo [118] [12] is an extension of MAMuJoCo [244]. In Safe MAMuJoCo, safety-aware agents have to learn not only the skillful manipulations of a robot, but also how to avoid crashing into unsafe obstacles and positions. In particular, the background environment, agents, physics simulator, and reward function are preserved. However, unlike its predecessor, a Safe MAMuJoCo environment comes with obstacles like walls or bombs. Furthermore, with the increasing risk of an agent stumbling upon an obstacle,

[12] https://github.com/chauncygu/Safe-Multi-Agent-Mujoco.git

the environment emits cost [47]. According to the scheme in [355], we characterize the cost functions for each task; the examples of Safe MAMuJoCo robots are shown in Figure 8 and the example tasks are shown in Figure 9.

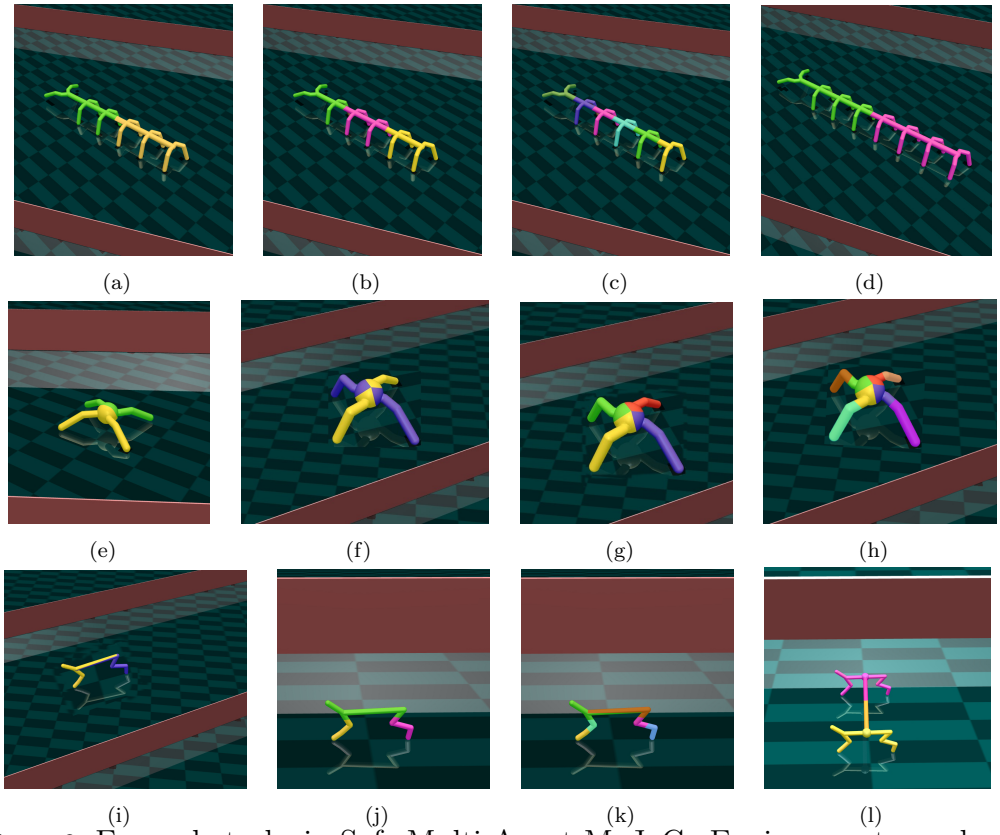
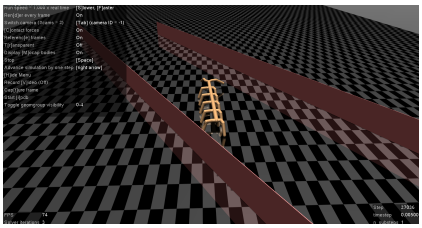


Figure 8: Example tasks in Safe Multi-Agent MuJoCo Environments, such as eight-agent Ant tasks and six-agent HalfCheetah tasks. Body parts of different colours are controlled by different agents. Agents jointly learn to manipulate the robot, while avoiding crashing into unsafe red areas, for details, see [118] (Adapt the figures with permission from [118]).

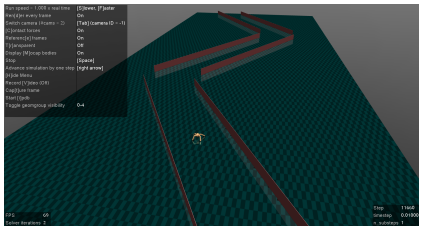
6.2.2. Safe Multi-Agent Robosuite

Safe Multi-Agent Robotsuite (Safe MARobosuite) [118] [13], shown in Figure 10, has been developed on the basis of Robosuite [367] which is a popular robotic arm benchmark for single-agent reinforcement learning. In

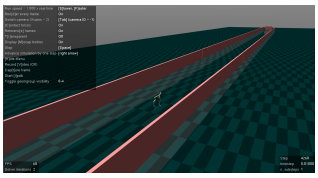
[13] https://github.com/chauncygu/Safe-Multi-Agent-Robosuite.git



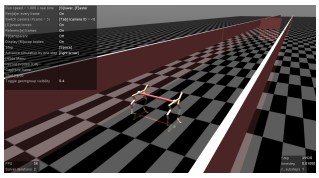
(a) ManyAgent Ant Task 1.0



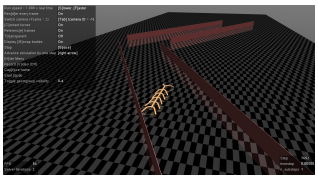
(b) Ant Task 1.0



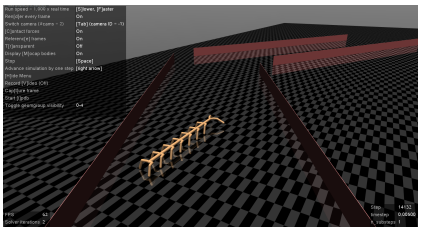
(c) HalfCheetah Task 1.0



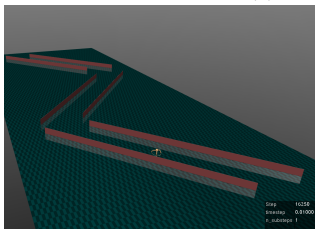
(d) Couple HalfCheetah Task 1.0



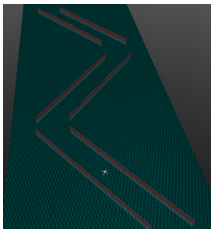
(e) ManyAgent Ant Task 2.1



(f) ManyAgent Ant Task 2.2



(g) Ant Task 2.1



(h) Ant Task 2.2

Figure 9: Specific tasks in Safe Multi-Agent MuJoCo Environments, e.g., 4x2-Ant Task 1.0, three folding jagged walls are Incorporated into the task; 3x2-ManyAgent Ant Task 2.1, two folding line walls is generated in the task. for details, see [118](Adapt the figures with permission from [118]).

Safe MARobosuite, multiple agents are set up according to the robot joint settings, and each agent controls every joint or several joints. A Lift task, for example, can be divided up into 2 agents (2x4 Lift), 4 agents (4x2 Lift), 8 agents (8x1 Lift); for Stack tasks, similarly, tasks can be provided with 2 agents (2x4 Stack), 4 agents (4x2 Stack), 8 agents (8x1 Stack); for a two-robot TwoArmLift task, it can be divided up into: 2 agents (2x8 TwoArmLift), 4 agents (4x4 TwoArmLift), 8 agents (8x2 TwoArmLift), 16 agents (16x1 TwoArmLift).

Moreover, Safe MARobosuite is also a fully cooperative, continuous, and decentralized benchmark that considers the constraints of robot safety. Where Franka robots are used to conduct each task, each agent can observe partial environmental information (such as the velocity and position). More importantly, Safe MARobosuite can be easily used for modular robots and make robots have good robustness and scalability. In many real-world applications, modular robots can be quickly paired and assembled for different tasks [10]. For instance, when communication bandwidth is limited, or some joints of robotic arms are broken, leading to causes malfunctioning communication, modular robots can still work well. The reward setting is the same as Robosuite [367], and the cost design is used to prevent robots from crashing into unsafe areas.

6.2.3. Safe Multi-Agent Isaac Gym

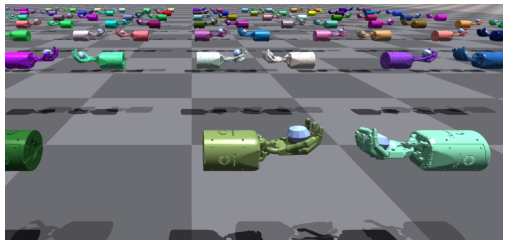
Safe Multi-Agent Isaac Gym (Safe MAIG) [14] is a high-performance environment that uses GPU for trajectory sampling and logical computation based on Isaac Gym [204], see Figure 11. The computation speed of Safe MAIG is almost ten times that of Safe MAuJoCo and Safe MARobosuite on the same server, in the same task, using the same algorithm. The communication speed is also better than Safe MAMuJoCo and Safe MARobosuite. But the memory might need to be optimized between CPU and GPU, since GPU memory is generally too small.

7. Challenges and Outlook of Safe Reinforcement Learning

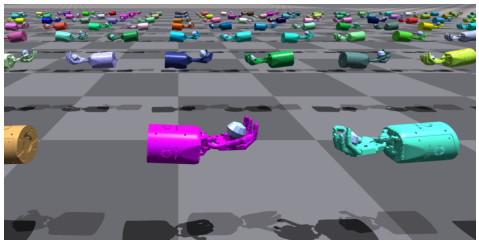
When deploying RL in real-world applications, numerous challenges emerge throughout the implementation process. In this section, the **"Safety Challenges"** problem is investigated by proposing several significant challenges we

[14] https://github.com/chauncygu/Safe-Multi-Agent-Isaac-Gym.git

Figure 10: Example tasks in Safe MARobosuite Environments, e.g., two-agent Lift tasks, eight-agent Lift tasks and fourteen-agent TwoArmPegInHole tasks. Agents jointly learn to manipulate the robot, while avoiding crashing into unsafe areas, for details, see [118](Adapt the figures with permission from [118]).

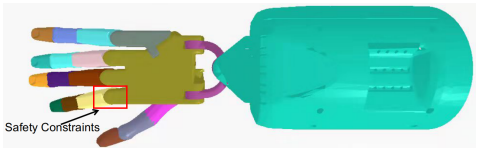


(a) ShadowHandOver 2x6



(b) ShadowHandCatchUnderarm 2x6

(c) ShadowHandReOrientation 4x3



(d) A hand showing safety constraints

Figure 11: Safe multi-agent Isaac Gym environments. Different robot body parts of different colours are manipulated by different agents in environments. Agents jointly learn how to control the robot, whilst the safety constraints are not violated (Adapt the figures with author permission).

need to address moving forward. Additionally, we identify potential research directions for enhancing the safety of RL systems. Except for the following challenges and research directions, we will investigate meta-safe RL [158] and robust RL [105, 239] in the future.

7.1. Human-Compatible Safe Reinforcement Learning

Modern safe RL algorithms rely on humans to state the objectives of safety. While humans understand the dangers, the potential risks are less noticed. Below we discuss two challenges in human preference statements and concerns on ethics and morality.

- **Human preference statement.** Many challenges are posed as AI agents are frequently evaluated in terms of performance measures, such as human-stated rewards. On the one hand, while it is usually assumed that humans are acting honestly in specifying their preferences, such as by rewards or demonstrations, the consequence of humans misstating their objectives is commonly underestimated. Humans may maliciously or unintentionally misstate their preferences, leading the safe RL agent to perform unexpected implementations. One example is the Tay chatbot from Microsoft; prankster users falsify their demonstrations and train Tay to mix racist comments into its dialogue [97]. On the other hand, multiple humans might be involved in training one safe RL agent. Thus, agents have to learn to strike a balance between the widely different human preferences. Earlier attempt [126] considers one agent vs one human scenario. However, many open questions remain, such as training robust agents against malicious users, personalizing assistance toward human preferences, etc.

- **Ethics and morality concerns.** In modern society, the human interrelationship is built based on social or moral norms. While reinforcement learning agents are deployed to the real world, they start having impacts on each other, turning into a multi-agent system, in which norms act similarly in human society on agents. Therefore, the decisions made by agents always involve ethical issues. For example, social dilemmas will emerge from the relation between individual goals and overall interests [141, 179]. The conflicts are produced when each agent aims to maximize its benefit. For another example, consider a *trolley problem* [99]. When the agent is faced with the choice of either harming multiple

people on the current track or one person by diverting the train, what would you expect the safe agents to choose? Or, more realistically, when the driving agent is about to bump into a lorry, it can swerve off the road to the left to save itself. However, there is a bike on the left. How could the driving agent make decisions? A human driver's knee-jerk reaction might be swerving left to save itself. However, the decision of the driving agent depends on its value systems. How to leverage the different value systems to enable safe agents to make ethical decisions is an open question.

7.2. Industrial Deployment Standards for Safe Reinforcement Learning

Although safe RL has been developed with a wealth of well-understood methods and algorithms in recent years [107, 159, 195], to our knowledge, there is no RL deployment standard for industrial applications, including technical standards and morality standards, etc. More attention is required on deployment standards and the alignment between academia and industries. Applications include robotics, autonomous driving, recommendation systems, finance, *et al.* We should tailor specific deployment standards to specific applications.

- **Technical standards.** In a technical standard, we need to think about how much efficiency RL can generate, how much time and money can be saved using RL methods, what environments can be handled with RL, how to design cost and reward functions considering the balance between RL reward, performance and safety, etc.

- **Law standards.** Human-machine interaction needs to be considered in legal judgments, for example, when robots hurt humans due to programming errors using RL methods. Furthermore, we need to determine how responsibilities are divided, e.g., do programmers of robots need to take more responsibility, or do robot users need to take more responsibility?

7.3. Safety Guarantee Efficiently in Large Number of Agents' Environments

Since the decision-making space is incredibly large when the number of agents increases in a safe MARL setting, it is not easy to optimize the multi-agent policy to finish tasks safely [350, 360]. Thus, efficiently guaranteeing safety in an environment with a large number of agents, e.g., 1 million agents, is still a challenging problem.

- **Theory of safe MARL.** Theory and experiments of safe MARL for massive swarm robots should be considered in the future, e.g., the convergence problem remains open for massive safe MARL methods without strong assumptions. Furthermore, optimizing sample complexity and stability within safe MARL environments is essential. Specifically, greater emphasis should be placed on the following key points regarding the theory of safe MARL. (1). Credit assignment both in cost and reward. In cooperative, competitive, and mixed game settings, it is crucial to contemplate the trade-off between reward and safety performance, e.g., a policy should be searched to minimize each agent's cost value while improving reward. Furthermore, it is necessary to optimize the precise cost value for each agent, and consider each agent's cost credit. (2). Nonstability. In a multi-agent system, when an agent takes actions, which will influence other agents' actions and may make other agents get worse reward value or unsafe. (3). Scalability. In safe MARL settings, when the number of agents becomes large, such as one billion agents, it will be challenging for computation and hard to ensure agents' safety, since it is almost impossible to estimate each agent's Q values or V value simultaneously.

- **Multiple heterogeneous constraints.** It still needs to be determined how multiple heterogeneous constraints should be handled in multi-agent systems for safe MARL. To our knowledge, almost no study has investigated multiple heterogeneous constraints for MARL. For example, when different agents encounter different heterogeneous constraints, we need to study how to balance different heterogeneous constraints to optimize different agents' policies for safety while improving reward values.

- **Carry out complex multi-agent tasks safely.** For instance, swarm robots perform the encirclement task, and then rescue the hostages from the dangerous scene safely. Currently, most safe MARL algorithms could have some challenging issues while conducting complex tasks, e.g., time-consuming issues or convergence issues.

- **Robust MARL.** One of the concerns in robust MARL settings: ensuring zero cost in different tasks without tuning parameters using the same algorithms is still open [133]. Another concern is when we transfer the safe MARL simulation results to real-world applications, it is still

unclear regarding how to handle the mismatch between the nominal and real system, since there may be the sensor noise [155] generated by fault sensors or disturbing sensor information transferred by adversary attacks.

- **Trade-off Balances.** The trade-off balance between exploration and exploitation is a dilemma in RL or MARL. Safe RL or safe MARL has the same problem. More particularly, there is another dilemma that is the trade-off between reward and cost, which is different from the exploration and exploitation, since each action can result in the change of reward and cost simultaneously, it is a multi-objective problem. Thus, in safe MARL settings, addressing the two balances mentioned earlier is imperative. In competitive game settings, efficiently modeling an opponent's decisions with limited information and implementing safe actions must be ascertained. In cooperative game settings, searching for a policy to guarantee the whole team's reward monotonic improvement while adhering to individual agent constraints is essential. In mixed-game settings, the optimization of both local and overall rewards while executing safe actions needs to be determined.

7.4. Possible Directions for Future Safe Reinforcement Learning Research

- Possible direction one: Safe MARL with game theory. How to solve the above challenges by leveraging game theory is a primary direction, since we can consider different games for real-world applications in different game settings, such as cooperative games or competitive games; how to optimize safety in an extensive form game is also helpful in real-world applications. For instance, In a fencing competition, we need to determine how to ensure that two agents complete the task while ensuring agents' safety in the game process.

- Possible direction two: Safe RL with information theory. Information theory may be useful to handle the uncertainty reward signal and cost estimation, and efficiently address the problem of large-scale MARL environments. For example, we can use information coding theory to construct the representation of different agent actions or reward signals.

- Possible direction three: Safe MARL with human-brain theory and biology inspiration theory. We can draw some inspiration from the laws of biology to design safe MARL algorithms. For example, we learn the

flying rules of geese, understand how geese form a certain formation, and ensure the safety of each geese.

- Possible direction four: Learn safe and diverse behaviors from human feedback, like ChatGPT[15]. During human-robot interaction, robots need to pay more effort into learning safe and diverse behaviors from humans rather than discriminatory and illegal behaviors. We envision that a robot needs to adapt to different tasks and learn to satisfy different people's preferences after a robot can learn safe behaviors from one person. In such a scenario, a robot needs to learn how to conduct safe behaviors with multiple persons. For example, feedback from humans can be used as training data to boost the relationship between robots and humans, and help to guarantee exploration safety [100]. Moreover, behavior diversity is a critical factor to successful multi-agent systems [349], and diverse behaviors can be leveraged to search for safe policies for safe robot learning.

- Possible direction five: Human-robot interactions. Learning from interactions with non-expert users is essential for long-term SRRL. For example, an earlier attempt [126] considers learning the user's reward signal from human-machine cooperation. The prerequisite that robots work with humans safely and efficiently is mutualism. Robots have to learn human preferences and comprehend human intentions and actions for successful collaboration with humans. By understanding what humans want and anticipate, robots can better adapt to their environment and receive enhanced support from their human counterparts. This mutual understanding fosters an effective and harmonious human-robot collaboration. Human behavior modeling [79] and realistic interactive modeling [368] can be the potential solutions for robots safely inheriting human preferences and learning more about humans.

8. Conclusion

We carefully review safe RL methods from the past 20 years, attempt to answer the key safe RL question around the investigation of safety research with **"2H3W"** problems, and provide a clear clue for further safe RL research.

[15] https://openai.com/blog/chatgpt/

Firstly, five critical safe RL problems are posed, and the model-based and model-free RL methods are analyzed in a unified safety framework. Secondly, the sample complexity and convergence of each method are investigated briefly. Thirdly, applications of safe RL are analyzed, for example, in the fields of autonomous driving and robotics. Fourthly, the benchmarks for safe RL communities are revealed, which may help RL take further toward real-world applications. Finally, the challenging problems that confront us during RL applications in safe RL domains are pointed out for future research.

Acknowledgments

This work was partially supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3). We would like to express our gratitude to Ming Jin, Yaodong Yang, and Hanna Krasowski for their invaluable suggestions during our safe RL research.

References

[1] Towards verifiable and safe model-free reinforcement learning. CEUR Workshop Proceedings, 2020.

[2] F Vazquez Abad, Vikram Krishnamurthy, Katerine Martin, and Irina Baltcheva. Self learning control of constrained markov chains-a gradient approach. In *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.*, volume 2, pages 1940–1945. IEEE, 2002.

[3] Felisa J Vázquez Abad and Vikram Krishnamurthy. Self learning control of constrained markov decision processes–a gradient approach. *Les Cahiers du GERAD ISSN*, 711:2440, 2003.

[4] Pieter Abbeel, Adam Coates, and Andrew Y Ng. Autonomous helicopter aerobatics through apprenticeship learning. *The International Journal of Robotics Research*, 29(13):1608–1639, 2010.

[5] Naoki Abe, Prem Melville, Cezar Pendus, Chandan K Reddy, David L Jensen, Vince P Thomas, James J Bennett, Gary F Anderson, Brent R Cooley, Melissa Kowalczyk, et al. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84, 2010.

[6] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of International Conference on Machine Learning (ICML)*, volume 70, pages 22–31, 2017.

[7] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[8] Anayo K Akametalu, Jaime F Fisac, Jeremy H Gillula, Shahab Kaynama, Melanie N Zeilinger, and Claire J Tomlin. Reachability-based safe learning with gaussian processes. In *53rd IEEE Conference on Decision and Control*, pages 1424–1431. IEEE, 2014.

[9] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[10] Matthias Althoff, Andrea Giusti, Stefan B Liu, and Aaron Pereira. Effortless creation of safe robots from modules through self-programming and self-verification. *Science Robotics*, 4(31), 2019.

[11] Eitan Altman. *Constrained Markov decision processes*. PhD thesis, INRIA, 1995.

[12] Eitan Altman. *Constrained Markov decision processes*. CRC Press, 1999.

[13] Aaron D Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *2019 18th European control conference (ECC)*, pages 3420–3431. IEEE, 2019.

[14] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[15] Greg Anderson, Abhinav Verma, Isil Dillig, and Swarat Chaudhuri. Neurosymbolic reinforcement learning with formally verified exploration. *Advances in neural information processing systems*, 33:6172–6183, 2020.

[16] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[17] Anil Aswani, Humberto Gonzalez, S Shankar Sastry, and Claire Tomlin. Provably safe and robust learning-based model predictive control. *Automatica*, 49(5):1216–1226, 2013.

[18] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: a

comprehensive overview and field guide for future research directions. *arXiv preprint arXiv:2112.11561*, 2021.

[19] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Towards safe, explainable, and regulated autonomous driving. *arXiv preprint arXiv:2111.10518*, 2021.

[20] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

[21] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[22] J Andrew Bagnell and Jeff G Schneider. Autonomous helicopter control using reinforcement learning policy search methods. In *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, volume 2, pages 1615–1620. IEEE, 2001.

[23] Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3682–3689, 2022.

[24] Olivier Bardou, Noufel Frikha, and Gilles Pages. Computing var and cvar using stochastic approximation and adaptive unconstrained importance sampling. 2009.

[25] Rafael Basso, Balázs Kulcsár, Ivan Sanchez-Diaz, and Xiaobo Qu. Dynamic stochastic electric vehicle routing with safe reinforcement learning. *Transportation research part E: logistics and transportation review*, 157:102496, 2022.

[26] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. *Advances in neural information processing systems*, 31, 2018.

[27] Arnab Basu, Tirthankar Bhattacharyya, and Vivek S Borkar. A learning algorithm for risk-sensitive cost. *Mathematics of operations research*, 33(4):880–898, 2008.

[28] Jared J Beard and Ali Baheri. Safety verification of autonomous systems: A multi-fidelity reinforcement learning approach. *arXiv preprint arXiv:2203.03451*, 2022.

[29] Felix Berkenkamp and Angela P Schoellig. Safe and robust learning control with gaussian processes. In *2015 European Control Conference (ECC)*, pages 2496–2501. IEEE, 2015.

[30] Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30, 2017.

[31] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

[32] Dimitri Bertsekas. *Convex optimization algorithms*. Athena Scientific, 2015.

[33] Frederick J Beutler and Keith W Ross. Optimal policies for controlled markov chains with a constraint. *Journal of mathematical analysis and applications*, 112(1):236–252, 1985.

[34] Frederick J Beutler and Keith W Ross. Time-average optimal constrained semi-markov decision processes. *Advances in Applied Probability*, 18(2):341–359, 1986.

[35] Homanga Bharadhwaj, Aviral Kumar, Nicholas Rhinehart, Sergey Levine, Florian Shkurti, and Animesh Garg. Conservative safety critics for exploration. In *International Conference on Learning Representations (ICLR)*, 2021.

[36] Shalabh Bhatnagar and K Lakshmanan. An online actor–critic algorithm with function approximation for constrained markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708, 2012.

[37] Lorenzo Bisi, Luca Sabbioni, Edoardo Vittori, Matteo Papini, and Marcello Restelli. Risk-averse trust region optimization for reward-volatility reduction. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4583–4589, 2020.

[38] David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

[39] Kenneth Bogert and Prashant Doshi. Multi-robot inverse reinforcement learning under occlusion with estimation of state transitions. *Artificial Intelligence*, 263:46–73, 2018.

[40] Vivek S Borkar. Q-learning for risk-sensitive control. *Mathematics of operations research*, 27(2):294–311, 2002.

[41] Vivek S Borkar. An actor-critic algorithm for constrained markov decision processes. *Systems & control letters*, 54(3):207–213, 2005.

[42] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[43] Andreas D Bovopoulos and Aurel A Lazar. The effect of delayed feedback information on network performance. *Annals of Operations Research*, 36(1):101–124, 1992.

[44] Justin Boyan and Michael Littman. Packet routing in dynamically changing networks: A reinforcement learning approach. *Advances in neural information processing systems*, 6, 1993.

[45] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[46] Kianté Brantley, Miroslav Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[47] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[48] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.

[49] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5, 2021.

[50] Jonas Buchli, Farbod Farshidian, Alexander Winkler, Timothy Sandy, and Markus Giftthaler. Optimal and learning control for autonomous robots. *arXiv preprint arXiv:1708.09342*, 2017.

[51] Mingyu Cai and Cristian-Ioan Vasile. Safe-critical modular deep reinforcement learning with temporal logic through gaussian processes and control barrier functions. *arXiv preprint arXiv:2109.02791*, 2021.

[52] Miguel Calvo-Fullana, Santiago Paternain, Luiz FO Chamon, and Alejandro Ribeiro. State augmented constrained reinforcement learning: Overcoming the limitations of learning with rewards. In *International Conference on Machine Learning*, 2021.

[53] Zhong Cao, Shaobing Xu, Xinyu Jiao, Huei Peng, and Diange Yang. Trustworthy safety improvement for autonomous driving using reinforcement learning. *Transportation research part C: emerging technologies*, 138:103656, 2022.

[54] Noe Casas. Deep deterministic policy gradient for urban traffic light control. *arXiv preprint arXiv:1703.09035*, 2017.

[55] Yi Chen, Jing Dong, and Zhaoran Wang. A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*, 2021.

[56] Yuxiao Chen, Andrew W Singletary, and Aaron D Ames. Density functions for guaranteed safety on robotic systems. In *2020 American Control Conference (ACC)*, pages 3199–3204. IEEE, 2020.

[57] Richard Cheng, Gábor Orosz, Richard M Murray, and Joel W Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3387–3395, 2019.

[58] Sandeep Chinchali, Pan Hu, Tianshu Chu, Manu Sharma, Manu Bansal, Rakesh Misra, Marco Pavone, and Sachin Katti. Cellular network traffic scheduling with deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[59] Jason Choi, Fernando Castañeda, Claire J Tomlin, and Koushil Sreenath. Reinforcement learning for safety-critical control under model uncertainty, using control lyapunov functions and control barrier functions. In *Robotics: Science and Systems (RSS)*, 2020.

[60] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

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

[62] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[63] Yinlam Chow, Ofir Nachum, Aleksandra Faust, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. Lyapunov-based safe policy optimization for continuous control. *ICML Workshop RL4RealLife*, 2019.

[64] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in neural information processing systems*, 28, 2015.

[65] Tianshu Chu, Jie Wang, Lara Codecà, and Zhaojian Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 21(3):1086–1095, 2019.

[66] Edmund M Clarke, Thomas A Henzinger, Helmut Veith, Roderick Bloem, et al. *Handbook of model checking*, volume 10. Springer, 2018.

[67] Jeffery A Clouse and Paul E Utgoff. A teaching method for reinforcement learning. In *Machine learning proceedings 1992*, pages 92–101. Elsevier, 1992.

[68] Max H Cohen and Calin Belta. Safe exploration in model-based reinforcement learning using control barrier functions. *Automatica*, 147:110684, 2023.

[69] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. 2016.

[70] Alexander I Cowen-Rivers, Daniel Palenicek, Vincent Moens, Mohammed Amin Abdullah, Aivar Sootla, Jun Wang, and Haitham Bou-Ammar. Samba: Safe model-based & active reinforcement learning. *Machine Learning*, pages 1–31, 2022.

[71] Robert H Crites and Andrew G Barto. Elevator group control using multiple reinforcement learning agents. *Machine learning*, 33(2):235–262, 1998.

[72] Felipe Leno Da Silva, Garrett Warnell, Anna Helena Reali Costa, and Peter Stone. Agents teaching agents: a survey on inter-agent transfer learning. *Autonomous Agents and Multi-Agent Systems*, 34(1):1–17, 2020.

[73] Bo Dai, Albert Shaw, Niao He, Lihong Li, and Le Song. Boosting the actor with dual critic. In *International Conference on Learning Representations*, 2018.

[74] Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

[75] Yann-Michaël De Hauwere, Peter Vrancx, and Ann Nowé. Learning multi-agent state space representations. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 715–722, 2010.

[76] Frits De Nijs, Erwin Walraven, Mathijs De Weerdt, and Matthijs Spaan. Constrained multiagent markov decision processes: A taxonomy of problems and algorithms. *Journal of Artificial Intelligence Research*, 70:955–1001, 2021.

[77] Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pages 465–472. Citeseer, 2011.

[78] Erick Delage and Shie Mannor. Percentile optimization in uncertain markov decision processes with application to efficient exploration. In *Proceedings of the 24th international conference on Machine learning*, pages 225–232, 2007.

[79] Ou Deng and Qun Jin. Policy-based reinforcement learning for assortative matching in human behavior modeling. *arXiv:2211.03936*, 2022.

[80] Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained mdps. *Advances in Neural Information Processing Systems*, 36, 2024.

[81] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

[82] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo R Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[83] Yunlong Dong, Xiuchuan Tang, and Ye Yuan. Principled reward shaping for reinforcement learning via lyapunov stability theory. *Neurocomputing*, 393:83–90, 2020.

[84] John Doyle. Robust and optimal control. In *Proceedings of 35th IEEE Conference on Decision and Control*, volume 2, pages 1595–1598. IEEE, 1996.

[85] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pages 1329–1338. PMLR, 2016.

[86] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

[87] Kyle Dunlap, Mark Mote, Kai Delsing, and Kerianne L Hobbs. Run time assured reinforcement learning for safe satellite docking. In *AIAA SCITECH 2022 Forum*, page 1853, 2022.

[88] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[89] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

[90] Mahmoud El Chamie, Yue Yu, and Behçet Açıkmeşe. Convex synthesis of randomized policies for controlled markov chains with density safety upper bound constraints. In *2016 American Control Conference (ACC)*, pages 6290–6295. IEEE, 2016.

[91] Ingy ElSayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. Safe multi-agent reinforcement learning via shielding. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 483–491, 2021.

[92] Yousef Emam, Paul Glotfelter, and Magnus Egerstedt. Robust barrier functions for a fully autonomous, remotely accessible swarm-robotics testbed. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3984–3990. IEEE, 2019.

[93] Yousef Emam, Gennaro Notomista, Paul Glotfelter, Zsolt Kira, and Magnus Egerstedt. Safe reinforcement learning using robust control barrier functions. *IEEE Robotics and Automation Letters*, 2022.

[94] Jiameng Fan and Wenchao Li. Safety-guided deep reinforcement learning via online gaussian process estimation. *arXiv preprint arXiv:1903.02526*, 2019.

[95] Alon Farchy, Samuel Barrett, Patrick MacAlpine, and Peter Stone. Humanoid robots learning to walk faster: From the real world to simulation and back. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 39–46, 2013.

[96] Eugene A Feinberg. Constrained discounted markov decision processes and hamiltonian cycles. *Mathematics of Operations Research*, 25(1):130–140, 2000.

[97] Arnaud Fickinger, Simon Zhuang, Dylan Hadfield-Menell, and Stuart Russell. Multi-Principal Assistance Games. 2020.

[98] Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

[99] Philippa Foot. The problem of abortion and the doctrine of the double effect. *Oxford review*, 5, 1967.

[100] Christopher Frye and Ilya Feige. Parenting: Safe reinforcement learning from human input. *arXiv preprint arXiv:1902.06766*, 2019.

[101] Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[102] Nathan Fulton and André Platzer. Verifiably safe off-model reinforcement learning. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 413–430. Springer, 2019.

[103] Zoltán Gábor, Zsolt Kalmár, and Csaba Szepesvári. Multi-criteria reinforcement learning. In *Proceedings of International Conference on Machine Learning (ICML)*, volume 98, pages 197–205, 1998.

[104] Aditya Gahlawat, Pan Zhao, Andrew Patterson, Naira Hovakimyan, and Evangelos Theodorou. L1-gp: L1 adaptive control with bayesian learning. In *Learning for Dynamics and Control*, pages 826–837. PMLR, 2020.

[105] Weinan Gao, Chao Deng, Yi Jiang, and Zhong-Ping Jiang. Resilient reinforcement learning and robust output regulation under denial-of-service attacks. *Automatica*, 142:110366, 2022.

[106] Javier Garcia and Fernando Fernández. Safe exploration of state and action spaces in reinforcement learning. *Journal of Artificial Intelligence Research*, 45:515–564, 2012.

[107] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[108] Javier García and Diogo Shafie. Teaching a humanoid robot to walk faster through safe reinforcement learning. *Engineering Applications of Artificial Intelligence*, 88:103360, 2020.

[109] Alborz Geramifard, Joshua Redding, and Jonathan P How. Intelligent cooperative control architecture: a framework for performance improvement using safe learning. *Journal of Intelligent & Robotic Systems*, 72(1):83–103, 2013.

[110] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[111] Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained rl with linear function approximation. *Advances in Neural Information Processing Systems*, 35:13303–13315, 2022.

[112] Peter G Gipps. A behavioural car-following model for computer simulation. *Transportation Research Part B: Methodological*, 15(2):105–111, 1981.

[113] Egor Gladin, Maksim Lavrik-Karmazin, Karina Zainullina, Varvara Rudenko, Alexander Gasnikov, and Martin Takac. Algorithm for

constrained markov decision process with linear convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 11506–11533. PMLR, 2023.

[114] David Gouaillier, Vincent Hugel, Pierre Blazevic, Chris Kilner, Jérôme Monceaux, Pascal Lafourcade, Brice Marnier, Julien Serre, and Bruno Maisonnier. Mechatronic design of nao humanoid. In *2009 IEEE international conference on robotics and automation*, pages 769–774. IEEE, 2009.

[115] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[116] Shangding Gu, Guang Chen, Lijun Zhang, Jing Hou, Yingbai Hu, and Alois Knoll. Constrained reinforcement learning for vehicle motion planning with topological reachability analysis. *Robotics*, 11(4), 2022.

[117] Shangding Gu, Dianye Huang, Muning Wen, Guang Chen, and Alois Knoll. Safe multiagent learning with soft constrained policy optimization in real robot control. *IEEE Transactions on Industrial Informatics*, 2024.

[118] Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois Knoll, and Yaodong Yang. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319:103905, 2023.

[119] Shangding Gu, Bilgehan Sel, Yuhao Ding, Lu Wang, Qingwei Lin, Ming Jin, and Alois Knoll. Balance reward and safety optimization for safe reinforcement learning: A perspective of gradient manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21099–21106, 2024.

[120] Shangding Gu, Chunhui Zhou, Yuanqiao Wen, Changshi Xiao, Zhe Du, and Liang Huang. Path search of unmanned surface vehicle based on topological location. *Navigation of China*, 42(02):52–58, 2019.

[121] Shangding Gu, Chunhui Zhou, Yuanqiao Wen, Changshi Xiao, and Alois Knoll. Motion planning for an unmanned surface vehicle with

wind and current effects. *Journal of Marine Science and Engineering*, 10(3):420, 2022.

[122] Shangding Gu, Chunhui Zhou, Yuanqiao Wen, Xi Zhong, Man Zhu, Changshi Xiao, and Zhe Du. A motion planning method for unmanned surface vehicle in restricted waters. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, 234(2):332–345, 2020.

[123] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE, 2017.

[124] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR, 2018.

[125] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, 2018.

[126] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[127] Minghao Han, Lixian Tian, Yuanand Zhang, Jun Wang, and Wei Pan. Reinforcement learning control of constrained dynamic systems with uniformly ultimate boundedness stability guarantee. *arXiv preprint arXiv:2011.06882*, 2020.

[128] Alexander Hans, Daniel Schneegaß, Anton Maximilian Schäfer, and Steffen Udluft. Safe exploration for reinforcement learning. In *ESANN*, pages 143–148. Citeseer, 2008.

[129] Mohammadhosein Hasanbeig, Alessandro Abate, and Daniel Kroening. Cautious reinforcement learning with logical constraints. In *Proc. the 19th Int. Conf. AAMAS*, pages 483–491, 2020.

[130] Aria HasanzadeZonuzy, Dileep Kalathil, and Srinivas Shakkottai. Model-based reinforcement learning for infinite-horizon discounted constrained markov decision processes. *IJCAI 2021*, 2021.

[131] Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted mdps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[132] Matthias Heger. Consideration of risk in reinforcement learning. In *Machine Learning Proceedings 1994*, pages 105–111. Elsevier, 1994.

[133] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

[134] Lukas Hewing, Juraj Kabzan, and Melanie N Zeilinger. Cautious model predictive control using gaussian process regression. *IEEE Transactions on Control Systems Technology*, 28(6):2736–2743, 2019.

[135] Tobias Holicki, Carsten W Scherer, and Sebastian Trimpe. Controller design via experimental exploration with robustness guarantees. *IEEE Control Systems Letters*, 5(2):641–646, 2020.

[136] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[137] Chen Hou and Qianchuan Zhao. Optimization of web service-based control system for balance between network traffic and delay. *IEEE Transactions on Automation Science and Engineering*, 15(3):1152–1162, 2017.

[138] NAIRA HOVAKIMYAN and CHENGYU CAO. L1 adaptive control theory: Guaranteed robustness with fast adaptation. *IEEE CONTROL SYSTEMS MAGAZINE*, 1066(033X/11), 2011.

[139] Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.

[140] Junyan Hu, Hanlin Niu, Joaquin Carrasco, Barry Lennox, and Farshad Arvin. Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning. *IEEE Transactions on Vehicular Technology*, 69(12):14413–14423, 2020.

[141] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31, 2018.

[142] Subin Huh and Insoon Yang. Safe reinforcement learning for probabilistic reachability and safety specifications: A lyapunov-based approach. *arXiv preprint arXiv:2002.10126*, 2020.

[143] Nathan Hunt, Nathan Fulton, Sara Magliacane, Trong Nghia Hoang, Subhro Das, and Armando Solar-Lezama. Verifiably safe exploration for end-to-end reinforcement learning. In *Proceedings of the 24th International Conference on Hybrid Systems: Computation and Control*, pages 1–11, 2021.

[144] Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

[145] David Isele, Alireza Nakhaei, and Kikuo Fujimura. Safe reinforcement learning on autonomous vehicles. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–6. IEEE, 2018.

[146] Nils Jansen, Bettina Könighofer, Sebastian Junges, AC Serban, and Roderick Bloem. Safe reinforcement learning using probabilistic shields. 2020.

[147] Ashkan B Jeddi, Nariman L Dehghani, and Abdollah Shafieezadeh. Lyapunov-based uncertainty-aware safe reinforcement learning. *arXiv preprint arXiv:2107.13944*, 2021.

[148] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020.

[149] Sebastian Junges, Nils Jansen, Christian Dehnert, Ufuk Topcu, and Joost-Pieter Katoen. Safety-constrained reinforcement learning for

mdps. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 130–146. Springer, 2016.

[150] Yoshinobu Kadota, Masami Kurano, and Masami Yasuda. Discounted markov decision processes with utility constraints. *Computers & Mathematics with Applications*, 51(2):279–284, 2006.

[151] Krishna C. Kalagarla, Rahul Jain, and Pierluigi Nuzzo. A sample-efficient algorithm for episodic finite-horizon mdp with constraints. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8030–8037, May 2021.

[152] Lodewijk Kallenberg. *Linear programming and finite Markovian control problems*, volume 148. 01 1983.

[153] R Kalman and J Bertram. Control system analysis and design via the second method of lyapunov:(i) continuous-time systems (ii) discrete time systems. *IRE Transactions on Automatic Control*, 4(3):112–112, 1959.

[154] Gabriel Kalweit, Maria Huegle, Moritz Werling, and Joschka Boedecker. Deep constrained q-learning. *arXiv preprint arXiv:2003.09398*, 2020.

[155] Danial Kamran, Tizian Engelgeh, Marvin Busch, Johannes Fischer, and Christoph Stiller. Minimizing safety interference for safe and comfortable automated driving with distributional reinforcement learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1236–1243. IEEE, 2021.

[156] Sanket Kamthe and Marc Deisenroth. Data-efficient reinforcement learning with probabilistic model predictive control. In *International conference on artificial intelligence and statistics*, pages 1701–1710. PMLR, 2018.

[157] Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8248–8254. IEEE, 2019.

[158] Vanshaj Khattar, Yuhao Ding, Bilgehan Sel, Javad Lavaei, and Ming Jin. A cmdp-within-online framework for meta-safe reinforcement learning.

In *The Eleventh International Conference on Learning Representations*, 2022.

[159] Youngmin Kim, Richard Allmendinger, and Manuel López-Ibáñez. Safe learning and optimization techniques: Towards a survey of the state of the art. In *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*, pages 123–139. Springer, 2020.

[160] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[161] J Kober and J Peters. Policy search for motor primitives in robotics. In *Twenty-Second Annual Conference on Neural Information Processing Systems (NIPS 2008)*, pages 849–856. Curran, 2009.

[162] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[163] Nate Kohl and Peter Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 3, pages 2619–2624. IEEE, 2004.

[164] Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. Learning-based model predictive control for safe exploration. In *Conference on Decision and Control (CDC)*, pages 6059–6066. IEEE, 2018.

[165] Rogier Koppejan and Shimon Whiteson. Neuroevolutionary reinforcement learning for generalized control of simulated helicopters. *Evolutionary intelligence*, 4(4):219–241, 2011.

[166] Alec Koppel, Kaiqing Zhang, Hao Zhu, and Tamer Başar. Projected stochastic primal-dual method for constrained online learning with kernels. *IEEE Transactions on Signal Processing*, 67(10):2528–2542, 2019.

[167] Iordanis Koutsopoulos and Leandros Tassiulas. Control and optimization meet the smart power grid: Scheduling of power demands for

optimal energy management. In *Proceedings of the 2nd International Conference on Energy-efficient Computing and Networking*, pages 41–50, 2011.

[168] Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125. IEEE, 2018.

[169] Hanna Krasowski, Xiao Wang, and Matthias Althoff. Safe reinforcement learning for autonomous lane changing using set-based prediction. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7. IEEE, 2020.

[170] Pavlo Krokhmal, Jonas Palmquist, and Stanislav Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, 4:43–68, 2002.

[171] Jakub Grudzien Kuba, Ruiqing Chen, Munning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11251*, 2021.

[172] Jakub Grudzien Kuba, Muning Wen, Yaodong Yang, Linghui Meng, Shangding Gu, Haifeng Zhang, David Henry Mguni, and Jun Wang. Settling the variance of multi-agent policy gradients. *arXiv preprint arXiv:2108.08612*, 2021.

[173] Abhishek Kumar and Rajneesh Sharma. Fuzzy lyapunov reinforcement learning for non linear systems. *ISA transactions*, 67:151–159, 2017.

[174] H. J. Kushner and J. Yang. Analysis of adaptive step-size sa algorithms for parameter tracking. In *IEEE Conference on Decision and Control*, pages 1403–1410, 1995.

[175] Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, ALT'12, page 320–334, 2012.

[176] Martin Lauer and Martin Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *In Proceedings of the Seventeenth International Conference on Machine Learning*. Citeseer, 2000.

[177] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning (ICML)*, pages 3703–3712, 2019.

[178] Dennis Lee, Haoran Tang, Jeffrey O Zhang, Huazhe Xu, Trevor Darrell, and Pieter Abbeel. Modular architecture for starcraft ii with deep reinforcement learning. In *Fourteenth Artificial Intelligence and Interactive Digital Entertainment Conference*, 2018.

[179] JZ Leibo, VF Zambaldi, M Lanctot, J Marecki, and T Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *AAMAS*, volume 16, pages 464–473. ACM, 2017.

[180] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

[181] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

[182] David L Leottau, Javier Ruiz-del Solar, and Robert Babuška. Decentralized reinforcement learning of robot behaviors. *Artificial Intelligence*, 256:130–159, 2018.

[183] G. Valiant Leslie. A theory of the learnable. 1984.

[184] Hepeng Li, Zhiqiang Wan, and Haibo He. Constrained ev charging scheduling based on safe deep reinforcement learning. *IEEE Transactions on Smart Grid*, 11(3):2427–2439, 2019.

[185] Tianjiao Li, Ziwei Guan, Shaofeng Zou, Tengyu Xu, Yingbin Liang, and Guanghui Lan. Faster algorithm and sharper analysis for constrained markov decision process. *Operations Research Letters*, 54:107107, 2024.

[186] Qingkai Liang, Fanyu Que, and Eytan Modiano. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480*, 2018.

[187] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *International Conference on Learning Representation (ICLR)*, 2016.

[188] An Liu, Vincent KN Lau, and Borna Kananian. Stochastic successive convex approximation for non-convex constrained stochastic optimization. *IEEE Transactions on Signal Processing*, 67(16):4189–4203, 2019.

[189] Chenyi Liu, Nan Geng, Vaneet Aggarwal, Tian Lan, Yuan Yang, and Mingwei Xu. Cmix: Deep multi-agent reinforcement learning with peak and average constraints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 157–173. Springer, 2021.

[190] Puze Liu, Davide Tateo, Haitham Bou Ammar, and Jan Peters. Robot reinforcement learning on the constraint manifold. In *Conference on Robot Learning*, pages 1357–1366. PMLR, 2022.

[191] Tao Liu, Ruida Zhou, Dileep Kalathil, P. R. Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[192] Yongshuai Liu, Jiaxin Ding, and Xin Liu. A constrained reinforcement learning based approach for network slicing. In *2020 IEEE 28th International Conference on Network Protocols (ICNP)*, pages 1–6. IEEE, 2020.

[193] Yongshuai Liu, Jiaxin Ding, and Xin Liu. Ipo: Interior-point policy optimization under constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4940–4947, 2020.

[194] Yongshuai Liu, Jiaxin Ding, and Xin Liu. Resource allocation method for network slicing using constrained reinforcement learning. In *2021 IFIP Networking Conference (IFIP Networking)*, pages 1–3. IEEE, 2021.

[195] Yongshuai Liu, Avishai Halev, and Xin Liu. Policy learning with constraints in model-free reinforcement learning: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021.

[196] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 2575–2582. IEEE, 2018.

[197] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[198] Songtao Lu, Kaiqing Zhang, Tianyi Chen, Tamer Basar, and Lior Horesh. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8767–8775, 2021.

[199] Xiaozhen Lu, Liang Xiao, Guohang Niu, Xiangyang Ji, and Qian Wang. Safe exploration in wireless security: A safe reinforcement learning algorithm with hierarchical structure. *IEEE Transactions on Information Forensics and Security*, 17:732–743, 2022.

[200] Björn Lütjens, Michael Everett, and Jonathan P How. Safe reinforcement learning with model uncertainty estimates. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8662–8668. IEEE, 2019.

[201] Christian Lutz, Felix Atmanspacher, Armin Hornung, and Maren Bennewitz. Nao walking down a ramp autonomously. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5169–5170. IEEE, 2012.

[202] Haitong Ma, Jianyu Chen, Shengbo Eben, Ziyu Lin, Yang Guan, Yangang Ren, and Sifa Zheng. Model-based constrained reinforcement learning using generalized control barrier function. In *2021 IEEE/RSJ*

International Conference on Intelligent Robots and Systems (IROS), pages 4552–4559. IEEE, 2021.

[203] Xiaobai Ma, Jiachen Li, Mykel J Kochenderfer, David Isele, and Kikuo Fujimura. Reinforcement learning for autonomous driving with latent state inference and spatial-temporal relationships. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6064–6071. IEEE, 2021.

[204] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu based physics simulation for robot learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[205] Amol Mandhane, Anton Zhernov, Maribeth Rauh, Chenjie Gu, Miaosen Wang, Flora Xue, Wendy Shang, Derek Pang, Rene Claus, Ching-Han Chiang, et al. Muzero with self-competition for rate control in vp9 video compression. *arXiv preprint arXiv:2202.06626*, 2022.

[206] Shie Mannor and John N Tsitsiklis. Mean-variance optimization in markov decision processes. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 177–184, 2011.

[207] Zahra Marvi and Bahare Kiumarsi. Safe off-policy reinforcement learning using barrier functions. In *2020 American Control Conference (ACC)*, pages 2176–2181. IEEE, 2020.

[208] Zahra Marvi and Bahare Kiumarsi. Safe reinforcement learning: A control barrier function optimization approach. *International Journal of Robust and Nonlinear Control*, 31(6):1923–1940, 2021.

[209] Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

[210] Karl Mason and Santiago Grijalva. A review of reinforcement learning for autonomous building energy management. *Computers & Electrical Engineering*, 78:300–312, 2019.

[211] Nicholas Mastronarde and Mihaela van der Schaar. Fast reinforcement learning for energy-efficient wireless communication. *IEEE Transactions on Signal Processing*, 59(12):6262–6266, 2011.

[212] David Q Mayne, James B Rawlings, Christopher V Rao, and Pierre OM Scokaert. Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814, 2000.

[213] David Q Mayne, María M Seron, and SV Raković. Robust model predictive control of constrained linear systems with bounded disturbances. *Automatica*, 41(2):219–224, 2005.

[214] Çetin Meriçli and Manuela Veloso. Biped walk learning through playback and corrective demonstration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 1594–1599, 2010.

[215] Olivier Michel. Cyberbotics ltd. webots™: professional mobile robot simulation. *International Journal of Advanced Robotic Systems*, 1(1):5, 2004.

[216] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

[217] Zoran Miljković, Marko Mitić, Mihailo Lazarević, and Bojan Babić. Neural network reinforcement learning for visual control of robot manipulators. *Expert Systems with Applications*, 40(5):1721–1736, 2013.

[218] Branka Mirchevska, Manuel Blum, Lawrence Louis, Joschka Boedecker, and Moritz Werling. Reinforcement learning for autonomous maneuvering in highway scenarios. In *Workshop for Driving Assistance Systems and Autonomous Driving*, pages 32–41, 2017.

[219] Branka Mirchevska, Christian Pek, Moritz Werling, Matthias Althoff, and Joschka Boedecker. High-level decision making for safe and reasonable autonomous lane changing using reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2156–2162. IEEE, 2018.

[220] Sobhan Miryoosefi, Kianté Brantley, Hal Daumé III, Miroslav Dudík, and Robert Schapire. Reinforcement learning with convex constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[221] Sobhan Miryoosefi and Chi Jin. A simple reward-free approach to constrained reinforcement learning. *arXiv preprint arXiv:2107.05216*, 2021.

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

[223] Teodor Mihai Moldovan and Pieter Abbeel. Risk aversion in markov decision processes via near optimal chernoff bounds. In *NIPS*, pages 3140–3148, 2012.

[224] Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in markov decision processes. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1451–1458, 2012.

[225] Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained mdps. In *International Conference on Machine Learning*, pages 25303–25336. PMLR, 2023.

[226] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[227] Anitha Murugesan, Mohammad Moghadamfalahi, and Arunabh Chattopadhyay. Formal methods assisted training of safe reinforcement learning agents. In *NASA Formal Methods: 11th International Symposium, NFM 2019, Houston, TX, USA, May 7–9, 2019, Proceedings 11*, pages 333–340. Springer, 2019.

[228] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019.

[229] Ofir Nachum and Bo Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

[230] Mitio Nagumo. Über die lage der integralkurven gewöhnlicher differentialgleichungen. *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series*, 24:551–559, 1942.

[231] Duy Nguyen-Tuong and Jan Peters. Model learning for robot control: a survey. *Cognitive processing*, 12(4):319–340, 2011.

[232] Behzad Nikbin, Mohammad Reza Ranjbar, B Shafiee Sarjaz, and Hamed Shah-Hosseini. Biped robot walking using particle swarm optimization. In *The 16th Online World Conference on Soft Computing in Industrial Applications (WSC16)*, 2011.

[233] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[234] Masahiro Ono, Marco Pavone, Yoshiaki Kuwata, and J Balaram. Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39(4):555–571, 2015.

[235] Stefan Oßwald, Attila Görög, Armin Hornung, and Maren Bennewitz. Autonomous climbing of spiral staircases with humanoids. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4844–4849. IEEE, 2011.

[236] Chris J Ostafew, Angela P Schoellig, and Timothy D Barfoot. Robust constrained learning-based nmpc enabling reliable mobile robot path tracking. *The International Journal of Robotics Research*, 35(13):1547–1563, 2016.

[237] Ioannis Panageas, Georgios Piliouras, and Xiao Wang. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. *Advances in Neural Information Processing Systems*, 32, 2019.

[238] Dimitra Panagou, Dušan M Stipanović, and Petros G Voulgaris. Distributed coordination control for multi-robot networks using lyapunov-like barrier functions. *IEEE Transactions on Automatic Control*, 61(3):617–632, 2015.

[239] Bo Pang and Zhong-Ping Jiang. Robust reinforcement learning: A case study in linear quadratic regulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9303–9311, 2021.

[240] Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *arXiv preprint arXiv:1911.09101*, 2019.

[241] Santiago Paternain, Luiz FO Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[242] Shashank Pathak, Luca Pulina, and Armando Tacchella. Verification and repair of control policies for safe reinforcement learning. *Applied Intelligence*, 48(4):886–908, 2018.

[243] Christian Pek, Peter Zahn, and Matthias Althoff. Verifying the safety of lane change maneuvers of self-driving vehicles based on formalized traffic rules. In *Proc. of the IEEE Intelligent Vehicles Symposium*, 2017.

[244] Bei Peng, Tabish Rashid, Christian A Schroeder de Witt, Pierre-Alexandre Kamienny, Philip HS Torr, Wendelin Böhmer, and Shimon Whiteson. Facmac: Factored multi-agent centralised policy gradients. *arXiv preprint arXiv:2003.06709*, 2020.

[245] Peng Peng, Ying Wen, Yaodong Yang, Quan Yuan, Zhenkun Tang, Haitao Long, and Jun Wang. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069*, 2017.

[246] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[247] Theodore J Perkins and Andrew G Barto. Lyapunov-constrained action sets for reinforcement learning. In *ICML*, volume 1, pages 409–416, 2001.

[248] Theodore J Perkins and Andrew G Barto. Lyapunov design for safe reinforcement learning. *Journal of Machine Learning Research*, 3(Dec):803–832, 2002.

[249] Jan Peters, Katharina Mülling, and Yasemin Altun. Relative entropy policy search. In *AAAI*, pages 1607–1612, 2010.

[250] Tu-Hoa Pham, Giovanni De Magistris, and Ryuki Tachibana. Optlayer-practical constrained optimization for deep reinforcement learning in the real world. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6236–6243. IEEE, 2018.

[251] Amir Pnueli. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pages 46–57. IEEE, 1977.

[252] David Pollard. Asymptopia: an exposition of statistical asymptotic theory. 2000. *URL http://www. stat. yale. edu/pollard/Books/Asymptopia*, 2000.

[253] Kyriakos Polymenakos, Alessandro Abate, and Stephen Roberts. Safe policy search using gaussian process models. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1565–1573, 2019.

[254] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988.

[255] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[256] Zengyi Qin, Yuxiao Chen, and Chuchu Fan. Density constrained reinforcement learning. In *International Conference on Machine Learning*, pages 8682–8692. PMLR, 2021.

[257] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 4295–4304. PMLR, 2018.

[258] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[259] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7:1, 2019.

[260] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking Safe Exploration in Deep Reinforcement Learning. 2019.

[261] Kevin Regan and Craig Boutilier. Regret-based reward elicitation for markov decision processes. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 444–451, 2009.

[262] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[263] Florian Richter, Ryan K Orosco, and Michael C Yip. Open-sourced reinforcement learning environments for surgical robotics. *arXiv preprint arXiv:1903.02090*, 2019.

[264] Joshua Riley, Radu Calinescu, Colin Paterson, Daniel Kudenko, and Alec Banks. Reinforcement learning with quantitative verification for assured multi-agent policies. In *13th International Conference on Agents and Artificial Intelligence*. York, 2021.

[265] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[266] Ugo Rosolia and Francesco Borrelli. Learning model predictive control for iterative tasks. a data-driven control framework. *IEEE Transactions on Automatic Control*, 63(7):1883–1896, 2017.

[267] Keith W Ross. Randomized and past-dependent policies for markov decision processes with multiple constraints. *Operations Research*, 37(3):474–477, 1989.

[268] Keith W Ross and Ravi Varadarajan. Markov decision processes with sample path constraints: the communicating case. *Operations Research*, 37(5):780–790, 1989.

[269] Keith W Ross and Ravi Varadarajan. Multichain markov decision processes with a sample path constraint: A decomposition approach. *Mathematics of Operations Research*, 16(1):195–207, 1991.

[270] Reazul Hasan Russel, Mouhacine Benosman, and Jeroen Van Baar. Robust constrained-mdps: Soft-constrained robust policy optimization under model uncertainty. *arXiv preprint arXiv:2010.04870*, 2020.

[271] Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.

[272] Manuel S Santos and John Rust. Convergence properties of policy iteration. *SIAM Journal on Control and Optimization*, 42(6):2094–2115, 2004.

[273] Shankar Sastry, Marc Bodson, and James F Bartram. Adaptive control: Stability, convergence, and robustness. *Acoustical Society of America Journal*, 88(1):588–589, 1990.

[274] Harsh Satija, Philip Amortila, and Joelle Pineau. Constrained markov decision processes via backward value functions. In *International Conference on Machine Learning (ICML)*, pages 8502–8511, 2020.

[275] Makoto Sato, Hajime Kimura, and Shibenobu Kobayashi. Td algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence*, 16(3):353–362, 2001.

[276] Thomas Savage, Dongda Zhang, Max Mowbray, and Ehecatl Antonio Del Río Chanona. Model-free safe reinforcement learning for chemical processes using gaussian processes. *IFAC-PapersOnLine*, 54(3):504–509, 2021.

[277] Jeff Schneider, Weng-Keen Wong, Andrew Moore, and Martin Riedmiller. Distributed value functions. 1999.

[278] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[279] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, pages 1889–1897, 2015.

[280] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *International Conference on Learning Representations (ICLR)*, 2016.

[281] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[282] Rodolphe Sepulchre, Mrdjan Jankovic, and Petar V Kokotovic. *Constructive nonlinear control*. Springer Science & Business Media, 2012.

[283] Guy Shani, David Heckerman, Ronen I Brafman, and Craig Boutilier. An mdp-based recommender system. *Journal of Machine Learning Research*, 6(9), 2005.

[284] Ziyad Sheebaelhamd, Konstantinos Zisis, Athina Nisioti, Dimitris Gkouletsos, Dario Pavllo, and Jonas Kohler. Safe deep reinforcement learning for multi-agent systems with continuous action spaces. *arXiv preprint arXiv:2108.03952*, 2021.

[285] Salomon Sickert, Javier Esparza, Stefan Jaax, and Jan Křetínskỳ. Limit-deterministic büchi automata for linear temporal logic. In *International Conference on Computer Aided Verification*, pages 312–332. Springer, 2016.

[286] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[287] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[288] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[289] Arambam James Singh, Akshat Kumar, and Hoong Chuin Lau. Hierarchical multiagent reinforcement learning for maritime traffic management. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1278–1286, 2020.

[290] Ashudeep Singh, Yoni Halpern, Nithum Thain, Konstantina Christakopoulou, EH Chi, Jilin Chen, and Alex Beutel. Building healthy recommendation sequences for everyone: A safe reinforcement learning approach. In *FAccTRec Workshop*, 2020.

[291] Avi Singh, Huihan Liu, Gaoyue Zhou, Albert Yu, Nicholas Rhinehart, and Sergey Levine. Parrot: Data-driven behavioral priors for reinforcement learning. In *International Conference on Learning Representations*, 2020.

[292] Dylan Slack, Yinlam Chow, Bo Dai, and Nevan Wichers. Safer: Data-efficient and safe reinforcement learning via skill acquisition. *arXiv preprint arXiv:2202.04849*, 2022.

[293] Morton Slater. Lagrange multipliers revisited: a contribution to nonlinear programming. 1950.

[294] Sungryull Sohn, Sungtae Lee, Jongwook Choi, Harm van Seijen, Mehdi Fatemi, and Honglak Lee. Shortest-path constrained reinforcement learning for sparse reward tasks. In *International Conference on Machine Learning*, 2021.

[295] Raffaele Soloperto, Matthias A Müller, Sebastian Trimpe, and Frank Allgöwer. Learning-based robust model predictive control with state-dependent uncertainty. *IFAC-PapersOnLine*, 51(20):442–447, 2018.

[296] Angus Stevenson. *Oxford dictionary of English*. Oxford University Press, USA, 2010.

[297] Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International conference on machine learning*, pages 997–1005. PMLR, 2015.

[298] Endre Süli and David F Mayers. *An introduction to numerical analysis*. Cambridge university press, 2003.

[299] Peng Sun, Xinghai Sun, Lei Han, Jiechao Xiong, Qing Wang, Bo Li, Yang Zheng, Ji Liu, Yongsheng Liu, Han Liu, et al. Tstarbots: Defeating the cheating level builtin ai in starcraft ii in the full game. *arXiv preprint arXiv:1809.07193*, 2018.

[300] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[301] Csaba Szepesvári. Constrained MDPs and the reward hypothesis. http://readingsml.blogspot.com/2020/03/constrained-mdps-and-reward-hypothesis.html, 2020. Access on January 21, 2023.

[302] Aviv Tamar, Dotan Di Castro, and Shie Mannor. Policy gradients with variance related risk criteria. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1651–1658, 2012.

[303] Aviv Tamar, Yonatan Glassner, and Shie Mannor. Policy gradients beyond expectations: Conditional value-at-risk. *arXiv preprint arXiv:1404.3862*, 2014.

[304] Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[305] Jie Tang, Arjun Singh, Nimbus Goehausen, and Pieter Abbeel. Parameterized maneuver learning for autonomous helicopter flight. In *2010 IEEE international conference on robotics and automation*, pages 1142–1148. IEEE, 2010.

[306] Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. In *International Conference on Learning Representations*, 2019.

[307] Gerald Tesauro. Td-gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation*, 6(2):215–219, 1994.

[308] Gerald Tesauro et al. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.

[309] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *arXiv preprint arXiv:1805.11074*, 2018.

[310] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. *International Conference on Learning Representation (ICLR)*, 2019.

[311] Garrett Thomas, Yuping Luo, and Tengyu Ma. Safe reinforcement learning by imagining the near future. *Advances in Neural Information Processing Systems*, 34, 2021.

[312] Andrea L Thomaz and Cynthia Breazeal. Reinforcement learning with human teachers: evidence of feedback and guidance with implications for learning performance. In *Proceedings of the 21st national conference on Artificial intelligence-Volume 1*, pages 1000–1005, 2006.

[313] A. M. Thompson and W. R. Cluett. Stochastic iterative dynamic programming: A monte carlo approach to dual control. *Automatica*, 41(5):767–778, 2005.

[314] Judith Jarvis Thomson. Killing, letting die, and the trolley problem. *The monist*, 59(2):204–217, 1976.

[315] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[316] Ernesto Torres and Leonardo Garrido. Automated generation of cpg-based locomotion for robot nao. In *Robot Soccer World Cup*, pages 461–471. Springer, 2011.

[317] Hoang-Dung Tran, Feiyang Cai, Manzanas Lopez Diego, Patrick Musau, Taylor T Johnson, and Xenofon Koutsoukos. Safety verification of cyber-physical systems with reinforcement learning control. *ACM Transactions on Embedded Computing Systems (TECS)*, 18(5s):1–22, 2019.

[318] John N Tsitsiklis and Benjamin Van Roy. Optimal stopping of markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Transactions on Automatic Control*, 44(10):1840–1851, 1999.

[319] Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite markov decision processes with gaussian processes. *Advances in Neural Information Processing Systems*, 29, 2016.

[320] Sharan Vaswani, Lin Yang, and Csaba Szepesvári. Near-optimal sample complexity bounds for constrained mdps. *Advances in Neural Information Processing Systems*, 35:3110–3122, 2022.

[321] Abhinav Verma, Hoang Le, Yisong Yue, and Swarat Chaudhuri. Imitation-projected programmatic reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[322] Alexander von Rohr, Matthias Neumann-Brosig, and Sebastian Trimpe. Probabilistic robust linear quadratic regulators with gaussian processes. In *Learning for Dynamics and Control*, pages 324–335. PMLR, 2021.

[323] Thanh Long Vu, Sayak Mukherjee, Renke Huang, and Qiuhua Huang. Barrier function-based safe reinforcement learning for emergency control of power systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3652–3657. IEEE, 2021.

[324] Quan Vuong, Yiming Zhang, and Keith W Ross. Supervised policy update for deep reinforcement learning. In *International Conference on Learning Representation (ICLR)*, 2019.

[325] Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806. PMLR, 2020.

[326] Akifumi Wachi, Yanan Sui, Yisong Yue, and Masahiro Ono. Safe exploration and optimization of constrained mdps using gaussian processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[327] Nolan Wagener, Byron Boots, and Ching-An Cheng. Safe reinforcement learning using advantage-based intervention. In *International Conference on Machine Learning*, 2021.

[328] Xiao Wang. Ensuring safety of learning-based motion planners using control barrier functions. *IEEE Robotics and Automation Letters*, 7(2):4773–4780, 2022.

[329] Xiaoyan Wang, Jun Peng, Shuqiu Li, and Bing Li. Formal reachability analysis for multi-agent reinforcement learning systems. *IEEE Access*, 9:45812–45821, 2021.

[330] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019.

[331] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

[332] CL Webots. Commercial mobile robot simulation software. *URL http://www. cyberbotics. com*, 2009.

[333] Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for constrained markov decision processes. *arXiv preprint arXiv:2106.01577*, 2021.

[334] Lu Wen, Jingliang Duan, Shengbo Eben Li, Shaobing Xu, and Huei Peng. Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7. IEEE, 2020.

[335] Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[336] Jim Woodcock, Peter Gorm Larsen, Juan Bicarregui, and John Fitzgerald. Formal methods: Practice and experience. *ACM computing surveys (CSUR)*, 41(4):1–36, 2009.

[337] Xiaoguang Wu, Shaowei Liu, Tianci Zhang, Lei Yang, Yanhui Li, and Tingjin Wang. Motion control for biped robot via ddpg-based deep reinforcement learning. In *2018 WRC Symposium on Advanced Robotics and Automation (WRC SARA)*, pages 40–45. IEEE, 2018.

[338] Liang Xiao, Yuzhen Ding, Jinhao Huang, Sicong Liu, Yuliang Tang, and Huaiyu Dai. Uav anti-jamming video transmissions with qoe guarantee: A reinforcement learning-based approach. *IEEE Transactions on Communications*, 69(9):5933–5947, 2021.

[339] Liang Xiao, Xiaozhen Lu, Tangwei Xu, Xiaoyue Wan, Wen Ji, and Yanyong Zhang. Reinforcement learning-based mobile offloading for edge computing against jamming and interference. *IEEE Transactions on Communications*, 68(10):6114–6126, 2020.

[340] Nuoya Xiong, Yihan Du, and Longbo Huang. Provably safe reinforcement learning with step-wise violation constraints. *Advances in Neural Information Processing Systems*, 36, 2024.

[341] Sijia Xu, Hongyu Kuang, Zhuang Zhi, Renjie Hu, Yang Liu, and Huyang Sun. Macro action selection with deep reinforcement learning in starcraft. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 94–99, 2019.

[342] Tengyu Xu, Yingbin Liang, and Guanghui Lan. A primal approach to constrained policy optimization: Global optimality and finite-time analysis. *arXiv preprint arXiv:2011.05869*, 2020.

[343] Tengyu Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, pages 11480–11491. PMLR, 2021.

[344] Long Yang, Jiaming Ji, Juntao Dai, Yu Zhang, Pengfei Li, and Gang Pan. Cup: A conservative update policy algorithm for safe reinforcement learning. *arXiv preprint arXiv:2202.07565*, 2022.

[345] Long Yang, Minhao Shi, Qian Zheng, Wenjia Meng, and Gang Pan. A unified approach for multi-step temporal-difference learning with eligibility traces in reinforcement learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2984–2990, 2018.

[346] Long Yang, Gang Zheng, Haotian Zhang, Yu Zhang, Qian Zheng, Jun Wen, and Gang Pan. Policy optimization with stochastic mirror descent. *arXiv preprint arXiv:1906.10462*, 2019.

[347] Long Yang, Qian Zheng, and Gang Pan. Sample complexity of policy gradient finding second-order stationary points. In *AAAI*, 2021.

[348] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representations (ICLR)*, 2020.

[349] Yaodong Yang, Jun Luo, Ying Wen, Oliver Slumbers, Daniel Graves, Haitham Bou Ammar, Jun Wang, and Matthew E Taylor. Diverse auto-curriculum is critical for successful real-world multiagent learning systems. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 51–56, 2021.

[350] Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 2020.

[351] Yujie Yang, Yuxuan Jiang, Yichen Liu, Jianyu Chen, and Shengbo Eben Li. Model-free safe reinforcement learning through neural barrier certificate. *IEEE Robotics and Automation Letters*, 8(3):1295–1302, 2023.

[352] Ming Yu, Zhuoran Yang, Mladen Kolar, and Zhaoran Wang. Convergent policy optimization for safe reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[353] Zhaocong Yuan, Adam W Hall, Siqi Zhou, Lukas Brunke, Melissa Greeff, Jacopo Panerati, and Angela P Schoellig. safe-control-gym: a unified benchmark suite for safe learning-based control and reinforcement learning. *arXiv preprint arXiv:2109.06325*, 2021.

[354] Seung-Kook Yun and Ambarish Goswami. Hardware experiments of humanoid robot safe fall using aldebaran nao. In *2012 IEEE International Conference on Robotics and Automation*, pages 71–78. IEEE, 2012.

[355] Moritz A. Zanger, Karam Daaboul, and J. Marius Zöllner. Safe continuous control with constrained model-based policy optimization, 2021.

[356] Mario Zanon and Sébastien Gros. Safe reinforcement learning using robust mpc. *IEEE Transactions on Automatic Control*, 66(8):3638–3652, 2020.

[357] Sihan Zeng, Thinh T Doan, and Justin Romberg. Finite-time complexity of online primal-dual natural actor-critic algorithm for constrained markov decision processes. *arXiv preprint arXiv:2110.11383*, 2021.

[358] Baohe Zhang, Yuan Zhang, Lilli Frison, Thomas Brox, and Joschka Bödecker. Constrained reinforcement learning with smoothed log barrier function. *arXiv preprint arXiv:2403.14508*, 2024.

[359] Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 33:10571–10583, 2020.

[360] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

[361] Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

[362] Xiangyu Zhao, Changsheng Gu, Haoshenglun Zhang, Xiwang Yang, Xiaobing Liu, Hui Liu, and Jiliang Tang. Dear: Deep reinforcement learning for online advertising impression in recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 750–758, 2021.

[363] Tianqi Zheng, Pengcheng You, and Enrique Mallada. Constrained reinforcement learning via dissipative saddle flow dynamics. In *2022*

56th Asilomar Conference on Signals, Systems, and Computers, pages 1362–1366. IEEE, 2022.

[364] Chunhui Zhou, Shangding Gu, Yuanqiao Wen, Zhe Du, Changshi Xiao, Liang Huang, and Man Zhu. Motion planning for an unmanned surface vehicle based on topological position maps. *Ocean Engineering*, 198:106798, 2020.

[365] Chunhui Zhou, Shangding Gu, Yuanqiao Wen, Zhe Du, Changshi Xiao, Liang Huang, and Man Zhu. The review unmanned surface vehicle path planning: Based on multi-modality constraint. *Ocean Engineering*, 200:107043, 2020.

[366] He Zhu, Zikang Xiong, Stephen Magill, and Suresh Jagannathan. An inductive synthesis framework for verifiable reinforcement learning. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 686–701, 2019.

[367] Yuke Zhu, Josiah Wong, Ajay Mandlekar, and Roberto Martín-Martín. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

[368] Zhengbang Zhu, Shenyu Zhang, Yuzheng Zhuang, Yuecheng Liu, Minghuan Liu, Liyuan Mao, Ziqing Gong, Weinan Zhang, Shixiong Kai, Qiang Gu, et al. Rita: Boost autonomous driving simulators with realistic interactive traffic flow. *arXiv:2211.03408*, 2022.