Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Survey Track

A Survey of Constraint Formulations in Safe Reinforcement Learning

Akifumi Wachi[1] , **Xun Shen**[2] , **Yanan Sui**[3]
[1]LY Corporation
[2]Osaka University
[3]Tsinghua University
akifumi.wachi@lycorp.co.jp, shenxun@eei.eng.osaka-u.ac.jp, ysui@tsinghua.edu.cn

Abstract

Safety is critical when applying reinforcement learning (RL) to real-world problems. As a result, safe RL has emerged as a fundamental and powerful paradigm for optimizing an agent's policy while incorporating notions of safety. A prevalent safe RL approach is based on a constrained criterion, which seeks to maximize the expected cumulative reward subject to specific safety constraints. Despite recent effort to enhance safety in RL, a systematic understanding of the field remains difficult. This challenge stems from the diversity of constraint representations and little exploration of their interrelations. To bridge this knowledge gap, we present a comprehensive review of representative constraint formulations, along with a curated selection of algorithms designed specifically for each formulation. In addition, we elucidate the theoretical underpinnings that reveal the mathematical mutual relations among common problem formulations. We conclude with a discussion of the current state and future directions of safe reinforcement learning research.

1 Introduction

Reinforcement learning (RL, [Sutton and Barto, 1998]) is a powerful paradigm for solving sequential decision-making problems through evaluation and improvement, which has made significant progress in many fields such as games [Mnih *et al.*, 2015; Silver *et al.*, 2016; Wurman *et al.*, 2022], robotics [Levine *et al.*, 2016], data center cooling [Li *et al.*, 2019], finance [Hambly *et al.*, 2023], recommendation systems [Afsar *et al.*, 2022], and healthcare [Yu *et al.*, 2021]. Recently, particular attention has been paid to RL for its use in the fine-tuning of large language models (LLMs) under the name of reinforcement learning from human feedback (RLHF, [Ouyang *et al.*, 2022]). RL is a general concept that can be employed in many domains and has been spreading to various disciplines as a useful paradigm.

When applying RL to real-world problems, *safety* is an essential requirement [Amodei *et al.*, 2016]. Thus, *safe RL* has been actively studied in recent years so that the benefits of RL are realized while minimizing negative safety issues. Promising areas for safe RL applications include robotics [Pham *et al.*, 2018], autonomous driving [Shalev-Shwartz *et al.*, 2016], healthcare [Jia *et al.*, 2020], and many others. An emerging and crucial use case of safe RL involves refining LLMs using RLHF to align with human preferences. Specifically, it is critical to prevent harmfulness or bias (e.g., toxicity, discrimination) while maintaining the helpfulness of the generated sentences. For example, Safe RLHF [Dai *et al.*, 2023] is a representative approach to balance such a tradeoff. *Safe RL* is an active area of research in artificial intelligence, with extensive theoretical and practical investigations aimed at developing RL systems that exhibit safe and reliable behavior.

Safe RL is inherently a broad concept with different formulations for the different aspects of real-world safety-critical problems. Garcıa and Fernández [2015] provided an eminent survey of safe RL and categorized its optimization criteria into four groups: 1) constrained criteron [Geibel, 2006], 2) worst-case criteron [Heger, 1994], 3) risk-sensitive criteron [Borkar, 2002], and 4) others (e.g., *r-squared, value-at-risk*). This paper focuses on the constrained criterion of safe RL as a growing and powerful group of methods to optimize policies under safety constraints.

A noteworthy aspect of safe reinforcement learning research based on constrained optimization criteria is the existence of diverse representations for safety constraints, with little analysis the interrelationships and theoretical connections among these various formulations. Safe RL research in the last decade has focused on developing new algorithms, i.e. *how to solve the problem* while pursuing performance. New algorithms have continuously been developed under various formulations, thereby making it increasingly challenging to stay abreast of advancements in the field. There are several recent survey papers on safe RL [Kim *et al.*, 2020; Brunke *et al.*, 2021; Liu *et al.*, 2021; Gu *et al.*, 2022; Zhao *et al.*, 2023b], but they also focus on methods rather than formulations. As formulation represents the initial phase in comprehending safe RL or implementing algorithms in practical scenarios, it becomes imperative for the community to comprehensively survey the existing literature and lay the groundwork for acquiring a systematic comprehension.

Our contributions. This paper provides a comprehensive survey focusing on constraint *formulations* in safe reinforcement learning and introduces representative algorithms for

each formulation. Furthermore, we discuss the relationships between various constraint formulations by defining three theoretical notions: *transformability*, *generalizability*, and *conservative approximation*. Specifically, we present theoretical results demonstrating that there exist two problems, termed **Identical or More General Safe RL (IoMG-SafeRL)** problems, into which other common problems can be either transformed or conservatively approximated. The main contribution of this paper is to bridge the gaps between the safe RL problems with appropriate algorithms by organizing existing research with a focus on constraint formulation.

2 Preliminaries

We consider safe RL problems modeled using a constrained Markov decision process (CMDP, [Altman, 1999]), which is formally represented as

$$\mathcal{M} \cup \mathcal{C} \coloneqq \underbrace{\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, H, r, \gamma_r, \rho \rangle}_{\text{Standard MDP } (\mathcal{M})} \cup \mathcal{C}, \tag{1}$$

where $\mathcal{S} \coloneqq \{s\}$ is a state space, $\mathcal{A} \coloneqq \{a\}$ is an action space, and $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition, where $\mathcal{P}(s' \mid s, a)$ is the probability of transition from state s to state s' when action a is taken. Note that $\Delta(X)$ denotes the probability simplex over the set X. In addition, $H \in \mathbb{Z}_+$ is the (fixed) finite length of each episode, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $\gamma_r \in [0, 1)$ is the discount factor for the reward, and $\rho \in \Delta(\mathcal{S})$ is the initial state distribution. A key difference from a standard MDP $\mathcal{M}$ lies in the existence of an additional tuple $\mathcal{C}$ to represent safety constraints, which will be referred to as a "constraint tuple" in the rest of this paper.[1]

A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a function to map from states to distribution over actions. Let Π denote a policy class. Given a policy $\pi \in \Pi$, the value function is defined as

$$V^\pi_{r,h}(s) \coloneqq \mathbb{E}_\pi \left[\sum_{h'=h}^{H} \gamma_r^{h'} r(s_{h'}, a_{h'}) \,\middle|\, s_h = s \right], \tag{2}$$

where the expectation $\mathbb{E}_\pi$ is taken over the random state-action sequence $\{(s_{h'}, a_{h'})\}_{h'=h}^{H}$ induced by the policy π and the CMDP $\mathcal{M} \cup \mathcal{C}$. Since the initial state s_0 is sampled from ρ, we slightly abuse the notation and define

$$V^\pi_r(\rho) \coloneqq \mathbb{E}_{s \sim \rho} \left[V^\pi_{r,0}(s) \right]. \tag{3}$$

Crucially, this paper deals with safe RL involving the "constrained" policy optimization problem. A policy must be within the feasible policy space $\widehat{\Pi} \subseteq \Pi$ that satisfies the safety constraint based on the given constraint tuple $\mathcal{C}$. Therefore, the optimal policy $\pi^\star : \mathcal{S} \to \Delta(\mathcal{A})$ is defined as

$$\pi^\star \coloneqq \arg\max_{\pi \in \widehat{\Pi}} V^\pi_r(\rho). \tag{4}$$

An overall sequence for solving safe RL problems based on a constrained criterion is illustrated in Figure 1, which consists of 1) problem formulation and 2) policy optimization. An interesting yet complicated point for understanding safe RL research is that there are several formulations for representing the constrained tuple $\mathcal{C}$ and the feasible policy space $\widehat{\Pi}$ depending on the problem settings or applications that researchers have in their minds. In the next section, we will review seven common safety constraint representations that have been well-studied in safe RL literature.

[1]Though this paper considers finite-horizon discounted CMDPs, key ideas can be extended to infinite and/or undiscounted cases.

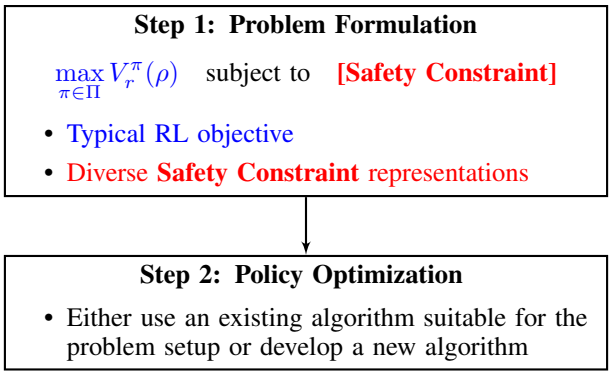

Figure 1: A typical sequence for solving safe RL problems based on constrained criteria. Due to the diversity of safety constraint representations and little discussion on their interrelations, it is not easy to understand safe RL research systematically. Unlike existing survey papers that focus on *methods*, we aim to provide a comprehensive survey from the perspective of *formulations* on safe RL.

3 Common Constraint Formulations

This paper deals with safe RL based on a constrained criterion. As presented shortly, for a CMDP $\mathcal{M} \cup \mathcal{C}$, common formulations discussed in this paper can be written as

$$\max_{\pi \in \Pi} V^\pi_r(\rho) \quad \text{s.t.} \quad f_\mathcal{C}(\pi) \leq 0. \tag{5}$$

In the rest of this paper, we call $f_\mathcal{C} : \Pi \to \mathbb{R}$ a *safety constraint function (SCF)* and let $\mathcal{F}_\mathcal{C} = \{f_\mathcal{C}\}$ denote a class of SCFs. Crucially, $\mathcal{C}$ contains the parametric information of the safety constraint; hence, $\mathcal{F}_\mathcal{C}$ is regarded as a class of SCFs that can be represented with $\mathcal{C}$. Note that, though $f_\mathcal{C}$ often depends on the parameters defined in the standard MDP $\mathcal{M}$ (e.g., $\mathcal{P}$), we omitted it from the notation for simplicity. The feasible policy space $\widehat{\Pi}$ depends on the structure and parameters of $f_\mathcal{C}$; hence is, we explicitly represent it as

$$\widehat{\Pi}(f_\mathcal{C}) \coloneqq \{\pi \in \Pi \mid f_\mathcal{C}(\pi) \leq 0\}. \tag{6}$$

In the following subsections, we will introduce common safety constraint representations and associated algorithms while organizing based on the aforementioned notations. Representative existing studies and algorithms are summarized according to the problem formulations in Table 1.

3.1 Expected Cumulative Safety Constraint

One of the most popular safe RL formulations is to represent the safety constraint using the same structure of the value function in terms of reward (i.e., $V^\pi_r(\rho)$). As with the reward function, we first define the value function for safety in a state

s at time h, denoted as

$$V^{\pi}_{c,h}(s) \coloneqq \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} \gamma_c^{h'} c(s_{h'}, a_{h'}) \,\middle|\, s_h = s\right].$$

where $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a safety cost function.[2] We then take expectation with respect to the initial state distribution ρ; that is, for a policy $\pi \in \Pi$, we have

$$V^{\pi}_{c}(\rho) \coloneqq \mathbb{E}_{s\sim\rho}\left[\, V^{\pi}_{c,0}(s)\,\right].$$

Therefore, we define the following safe RL problem based on the value function regarding the safety cost function.

Problem 1. Let a constraint tuple be $\mathcal{C} \coloneqq \langle\, c, \gamma_c, \xi \,\rangle$, where $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the safety cost function, $\gamma_c \in [0, 1)$ is the discount factor for the safety, and $\xi \in \mathbb{R}_+$ is the safety threshold. Then,

$$\max_{\pi} V^{\pi}_{r}(\rho) \quad \text{s.t.} \quad V^{\pi}_{c}(\rho) \le \xi. \tag{7}$$

Remark 1. In Problem 1, the SCF $f_{\mathcal{C}} : \Pi \to \mathbb{R}$ is defined as

$$f_{\mathcal{C}}(\pi) \coloneqq V^{\pi}_{c}(\rho) - \xi.$$

A reason why Problem 1 is popular is that the safety constraint has a high affinity with the value function in terms of reward. For example, when $\gamma_r = \gamma_c$, for any $\lambda \in \mathbb{R}$, we have

$$V^{\pi}_{r+\lambda c}(\rho) = V^{\pi}_{r}(\rho) + \lambda \cdot V^{\pi}_{c}(\rho),$$

which leads to so-called Lagrangian-based methods [Altman, 1999]. Thus, since the theoretical properties of the problem have been studied, several algorithms with theoretical guarantees on optimality or safety have been proposed [Ding *et al.*, 2021; Bura *et al.*, 2022]. Practically, many well-known algorithms, such as constrained policy optimization (CPO, [Achiam *et al.*, 2017]) or reward-constrained policy optimization (RCPO, [Tessler *et al.*, 2019]), are built upon Problem 1.

3.2 State Constraint

Another popular formulation involves leveraging the "state" constraints so that an agent avoids visiting a set of unsafe states, which is well-suited for the case where an autonomous robot needs to behave safely in a hazardous environment (e.g., a disaster site). This type of formulation has been widely adopted by previous studies on safe-critical robotics tasks [Thananjeyan *et al.*, 2021; Thomas *et al.*, 2021; Turchetta *et al.*, 2020], which is written as follows:

Problem 2. Let a constraint tuple be $\mathcal{C} \coloneqq \langle\, S_{\text{unsafe}}, \gamma_c, \xi \,\rangle$, where $S_{\text{unsafe}} \subseteq \mathcal{S}$ is the set of unsafe states, $\gamma_c \in [0, 1)$ is the discount factor for the safety cost function, and $\xi \in \mathbb{R}_+$ is the safety threshold. Then,

$$\max_{\pi} V^{\pi}_{r}(\rho) \quad \text{s.t.} \quad \mathbb{E}_{\pi}\left[\sum_{h=0}^{H} \gamma_c^h \mathbb{I}(s_h \in S_{\text{unsafe}})\right] \le \xi,$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Remark 2. In Problem 2, the SCF $f_{\mathcal{C}} : \Pi \to \mathbb{R}$ is defined as

$$f_{\mathcal{C}}(\pi) \coloneqq \mathbb{E}_{\pi}\left[\sum_{h=0}^{H} \gamma_c^h \mathbb{I}(s_h \in S_{\text{unsafe}})\right] - \xi.$$

[2] Without loss of generality, we assume that the safety cost function c is bounded between 0 and 1 for all (s, a) as with the reward function r. Especially, the assumption that the safety cost is *positive* is important for our theoretical analyses.

3.3 Joint Chance Constraint

Policy optimization under joint chance constraints has been studied especially in the field of control theory such as Ono *et al.* [2015] and Pfrommer *et al.* [2022]. This type of safety-constrained policy optimization problem is typically formulated as follows:

Problem 3. Let a constraint tuple be $\mathcal{C} \coloneqq \langle\, S_{\text{unsafe}}, \xi \,\rangle$, where $S_{\text{unsafe}} \subseteq \mathcal{S}$ is the set of unsafe states and $\xi \in \mathbb{R}_+$ is the safety threshold. Then,

$$\max_{\pi} V^{\pi}_{r}(\rho) \quad \text{s.t.} \quad \mathbb{P}_{\pi}\left[\bigvee_{h=0}^{H} s_h \in S_{\text{unsafe}}\right] \le \xi,$$

where $\mathbb{P}_{\pi}$ represents a probability that is computed over the random state-action sequences $\{(s_{h'}, a_{h'})\}_{h'=h}^{H}$ induced by the policy π and the CMDP $\mathcal{M} \cup \mathcal{C}$.

Remark 3. In Problem 3, the SCF $f_{\mathcal{C}} : \Pi \to \mathbb{R}$ is defined as

$$f_{\mathcal{C}}(\pi) \coloneqq \mathbb{P}_{\pi}\left[\bigvee_{h=0}^{H} s_h \in S_{\text{unsafe}}\right] - \xi.$$

It is quite challenging to directly solve the Problem 3 characterized by joint chance constraints; thus, most of the previous work does not directly deal with this type of constraint and uses some approximations or assumptions. For example, Pfrommer *et al.* [2022] assumes a known linear time-invariant dynamics. Also, Ono *et al.* [2015] conservatively approximate the joint chance constraint into the constraint with an additive structure as in Problem 2 using the following inequality (for more details, see the proof of Theorem 2):

$$\mathbb{P}_{\pi}\left[\bigvee_{h=1}^{H} s_h \in S_{\text{unsafe}}\right] \le \mathbb{E}_{\pi}\left[\sum_{h=1}^{H} \mathbb{I}(s_h \in S_{\text{unsafe}})\right].$$

3.4 Expected Instantaneous Safety Constraint with Time-variant Threshold

Though Problems 1, 2, and 3 formulate the long-term safety constraint with additive structures or joint chance constraints, there are a few papers (e.g., Pham *et al.* [2018]) that focus on "instantaneous" safety constraint. In this problem, an agent is required to satisfy a safety constraint at every time step.

Problem 4. Let a constraint tuple be $\mathcal{C} \coloneqq \langle\, c, \{\xi_h\}_{h=0}^{H} \,\rangle$, where $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the safety cost function, and $\xi_h \in \mathbb{R}_+$ is the safety threshold for time step $h \in [H]$. Then,

$$\max_{\pi} V^{\pi}_{r}(\rho) \quad \text{s.t.} \quad \mathbb{E}_{\pi}\left[\, c(s_h, a_h)\,\right] \le \xi_h,\ \forall h \in [H].$$

Remark 4. In Problem 4, the SCF $f_{\mathcal{C}} : \Pi \to \mathbb{R}$ is defined as

$$f_{\mathcal{C}}(\pi) \coloneqq \max_{h\in[H]} \Big(\mathbb{E}_{\pi}\left[\, c(s_h, a_h)\,\right] - \xi_h\Big).$$

3.5 Almost Surely Cumulative Safety Constraint

Unlike Problem 1 where the expectation $\mathbb{E}_{\pi}$ regarding the safety constraint is taken, we often want to guarantee safety *almost surely* (i.e., probability of 1). This problem has been recently studied [Sootla *et al.*, 2022a; Sootla *et al.*, 2022b], which is based on a stricter safety notion. In such cases, the safe RL problem is formulated as follows:

Problem 5. Let a constraint tuple be $\mathcal{C} \coloneqq \langle\, c, \gamma_c, \xi \,\rangle$, where $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the safety cost function, $\gamma_c \in [0, 1)$ is the discount factor for the safety, and $\xi \in \mathbb{R}_+$ is the safety threshold. Then,

$$\max_\pi V_r^\pi(\rho) \quad \text{s.t.} \quad \mathbb{P}_\pi \left[\sum_{h=0}^{H} \gamma_c^h c(s_h, a_h) \leq \xi \right] = 1.$$

Remark 5. In Problem 5, the SCF $f_\mathcal{C} : \Pi \to \mathbb{R}$ is defined as

$$f_\mathcal{C}(\pi) \coloneqq 1 - \mathbb{P}_\pi \left[\sum_{h=0}^{H} \gamma_c^h c(s_h, a_h) \leq \xi \right].$$

3.6 Almost Surely Instantaneous Safety Constraint with Time-invariant Threshold

Some existing studies formulate safe RL problems via an instantaneous constraint, attempting to ensure safe exploration [Sui *et al.*, 2015] even during learning while aiming for extremely safety-critical RL applications such as planetary exploration and medical treatment [Turchetta *et al.*, 2016; Wachi *et al.*, 2018; Wachi and Sui, 2020; Wang *et al.*, 2023]. Such studies require the agent to satisfy the following instantaneous safety constraint at every time step.

Problem 6. Let a constraint tuple be $\mathcal{C} \coloneqq \langle\, c, \xi \,\rangle$, where $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the safety cost function, and $\xi \in \mathbb{R}_+$ is the threshold. Then,

$$\max_\pi V_r^\pi(\rho) \quad \text{s.t.} \quad \mathbb{P}_\pi \left[\, c(s_h, a_h) \leq \xi \,\right] = 1, \ \forall h \in [H].$$

Remark 6. In Problem 6, the SCF $f_\mathcal{C} : \Pi \to \mathbb{R}$ is defined as

$$f_\mathcal{C}(\pi) \coloneqq 1 - \prod_{h=0}^{H} \mathbb{P}_\pi \left[\, c(s_h, a_h) \leq \xi \,\right].$$

This formulation is also related to notions in the control theory, called control barrier functions [Ames *et al.*, 2016; Cheng *et al.*, 2019] or Lyapunov functions [Berkenkamp *et al.*, 2017]. Note that while these functions are typically required to be positive, it is possible to use the same formulation in Problem 6 by defining them as $-c$ and setting $\xi = 0$.

3.7 Almost Surely Instantaneous Safety Constraint with Time-variant Threshold

As a similar formulation to Problem 6, Wachi *et al.* [2023] has recently introduced a problem called the generalized safe exploration (GSE) problem, which is written as follows:

Problem 7. Let a constraint tuple be $\mathcal{C} \coloneqq \langle\, c, \{\xi_h\}_{h=0}^H \,\rangle$, where $c : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the safety cost function, and $\xi_h \in \mathbb{R}_+$ is the safety threshold for time step $h \in [H]$. Then,

$$\max_\pi V_r^\pi(\rho) \quad \text{s.t.} \quad \mathbb{P}_\pi \left[\, c(s_h, a_h) \leq \xi_h \,\right] = 1, \forall h \in [H].$$

Remark 7. In Problem 7, the SCF $f_\mathcal{C} : \Pi \to \mathbb{R}$ is defined as

$$f_\mathcal{C}(\pi) \coloneqq 1 - \prod_{h=0}^{H} \mathbb{P}_\pi \left[\, c(s_h, a_h) \leq \xi_h \,\right].$$

This formulation is quite similar to Problem 6 with the only difference being that the safety threshold ξ_h is time-variant. An apparent benefit of Problem 7 compared to Problem 6 is that we can cover a wider range of applications such that the speed limit changes during driving. Additionally, it goes beyond that and offers us more important advantages regarding the theoretical relations with Problem 5, which we will present shortly in Theorem 3.

3.8 Other Constrained Formulations

In the line of safe RL work based on constraints, various attempts were left unexplored and not integrated into our theoretical framework within this paper.

A notable instance of such problem formulations defines a safety constraint using the *variance* of the return [Tamar *et al.*, 2012] which is represented as

$$\max_\pi V_r^\pi(\rho) \quad \text{s.t.} \quad \text{Var}\left[\, V_r^\pi(\rho) \,\right] < \xi.$$

The *variance* of the return is related to the Sharpe ratio [Sharpe, 1966] – the ratio between the expected profit and its standard deviation. Thus, this formulation is particularly useful in financial appliactions [Meng and Khushi, 2019].

Also, conventional RL algorithms depend on *ergodicity* assumption; that is, any state s is eventually reachable from any other state s' by following a suitable policy. This assumption does not hold in many practical applications since an agent cannot recover on its own after catastrophic actions. Moldovan and Abbeel [2012] removed the ergodicity assumption and proposed an algorithm in which an agent is required to guarantee returnability to the initial safe state.

While this paper has dealt with *numerical* safety constraints, there have been attempts to represent them using a certain language. For example, Fulton and Platzer [2018] or Hasanbeig *et al.* [2020] represent the safety constraint via formal languages such as linear temporal logic. This constraint representation is quite useful for leveraging human knowledge in RL, which leads to powerful solutions such as shielding methods [Alshiekh *et al.*, 2018; Li and Bastani, 2020]. Finally, Yang *et al.* [2021] uses natural language to represent safety constraints. Natural language is one of the most powerful mediums yet friendly representations to the general public. Given the recent remarkable progress and utilities of LLMs, this approach would be promising for applying safe RL to real-world AI systems.

4 Theoretical Relations among Common Constraint Formulations of Safe RL

In this section, we provide theoretical understandings regarding the interrelations among common safe RL formulations presented in Section 3.

4.1 Definitions

We first introduce and define two important notions called *transformability* and *generalizability*. Also, we define a notion called *conservative approximation*.

Definition 1 (Transformability)**.** Let $\mathcal{M} \cup \mathcal{C}_1$ and $\mathcal{M} \cup \mathcal{C}_2$ denote two CMDPs that are respectively characterized by different constraint tuples $\mathcal{C}_1$ and $\mathcal{C}_2$. Let $\mathcal{F}_{\mathcal{C}_1}$ and $\mathcal{F}_{\mathcal{C}_2}$ denote two

Problem	Type	Representative Work	Algorithm	Theoretical Guarantee		Open Source Software (OSS)
				Optimality	Safety	
Problem 1	Online	[Achiam *et al.*, 2017]	CPO	–	–	A, SSA, FSRL, SafePO, OmniSafe
		[Ray *et al.*, 2019]	TRPO-Lagrangian	–	–	A, SSA, FSRL, SafePO, OmniSafe
			PPO-Lagrangian	–	–	A, SSA, FSRL, SafePO, OmniSafe
		[Tessler *et al.*, 2019]	RCPO	–	–	A, SafePO, OmniSafe
		[Liu *et al.*, 2020]	IPO	–	–	A, OmniSafe
		[Yang *et al.*, 2020]	PCPO	–	–	A, SafePO, OmniSafe
		[Stooke *et al.*, 2020]	PID-Lagrangian	–	–	A, SafePO, OmniSafe
		[Zhang *et al.*, 2020]	FOCOPS	–	–	A, FSRL, SafePO, OmniSafe
		[Ding *et al.*, 2020]	NPG-PD	Y	C	–
		[Bharadhwaj *et al.*, 2021]	CSC	–	–	A
		[Ding *et al.*, 2021]	OPDOP	Y	C	–
		[Bai *et al.*, 2022]	CSPDA	Y	C	–
		[As *et al.*, 2021]	LAMBDA	–	–	A
		[Xu *et al.*, 2021]	CRPO	Y	C	OmniSafe
		[Yu *et al.*, 2022]	SEditor	–	–	A
		[Bura *et al.*, 2022]	DOPE	Y	T and C	–
		[Liu *et al.*, 2022]	CVPO	Y	C	A, FSRL
		[Zhang *et al.*, 2022]	P3O	–	–	A, OmniSafe
	Offline	[Le *et al.*, 2019]	CBPL	–	T and C	A
		[Lee *et al.*, 2021]	COptiDICE	–	T	A, OSRL, OmniSafe
		[Wu *et al.*, 2021]	CMOMDPs	Y	T and C	–
		[Xu *et al.*, 2022]	CPQ	–	T	A, OSRL
		[Liu *et al.*, 2023b]	CDT	–	T	A, OSRL
Problem 2	Online	[Turchetta *et al.*, 2020]	CISR	–	–	A
		[Thomas *et al.*, 2021]	SMBPO	–	C	A
		[Thananjeyan *et al.*, 2021]	Recovery RL	–	–	A
		[Wang *et al.*, 2023]	–	–	T and C	A
Problem 3	Online	[Ono *et al.*, 2015]	CCDP	–	T and C	–
		[Pfrommer *et al.*, 2022]	–	Y	T and C	–
		[Mowbray *et al.*, 2022]	–	–	T and C	A
		[Kordabad *et al.*, 2022]	–	–	T and C	–
Problem 4	Online	[Pham *et al.*, 2018]	OptLayer	–	T and C	A
		[Amani *et al.*, 2021]	SLUCB	Y	T and C	–
		[Zhao *et al.*, 2023a]	SCPO	Y	C	–
	Offline	[Amani and Yang, 2022]	Safe-DPVI	Y	T and C	–
Problem 5	Online	[Sootla *et al.*, 2022b]	Sauté RL	Y	C	A, SafePO, OmniSafe
		[Sootla *et al.*, 2022a]	Simmer RL	Y	C	A, SafePO, OmniSafe
Problem 6	Online	[Turchetta *et al.*, 2016]	SafeMDP	–	T and C	A
		[Berkenkamp *et al.*, 2017]	SMbRL	–	T and C	A
		[Fisac *et al.*, 2018]	–	–	T and C	–
		[Wachi *et al.*, 2018]	SafeExpOpt-MDP	–	T and C	A
		[Dalal *et al.*, 2018]	SafeLayer	–	T and C	A
		[Cheng *et al.*, 2019]	RL-CBF	–	T and C	A
		[Wachi and Sui, 2020]	SNO-MDP	Y	T and C	A
		[Wang *et al.*, 2023]	–	–	C	–
Problem 7	Online	[Shi *et al.*, 2023]	LSVI-NEW	Y	T and C	–
		[Wachi *et al.*, 2023]	MASE	Y	T and C	–

Table 1: Common safe RL formulations based on the constrained criterion and associated representative work. Type indicates whether each safety RL is based on online or offline RL settings. In the Theoretical Guarantee column, **Y** indicates the (near-)optimality of the policy obtained by an algorithm. Also, **T** means that safety is guaranteed during training, and **C** means that safety is guaranteed after convergence. Note that offline algorithms are inherently safe during training since there is no interaction between the agent and the environment. In the OSS column, **A** means a public authors' implementation exists, and **SSA** is an abbreviation of the Safety Starter Agent repository (Ray *et al.* [2019], https://github.com/openai/safety-starter-agents). Also, **FSRL** (Liu *et al.* [2023a], https://github.com/liuzuxin/FSRL), **OSRL** (Liu *et al.* [2023a], https://github.com/liuzuxin/OSRL), **SafePO** (Ji *et al.* [2023], https://github.com/PKU-Alignment/Safe-Policy-Optimization), and **OmniSafe** (Ji *et al.* [2023], https://github.com/PKU-Alignment/omnisafe) are recent and actively maintained repositories for online and offline safe RL, which will lead to the ease of the process of adopting safe RL algorithms.

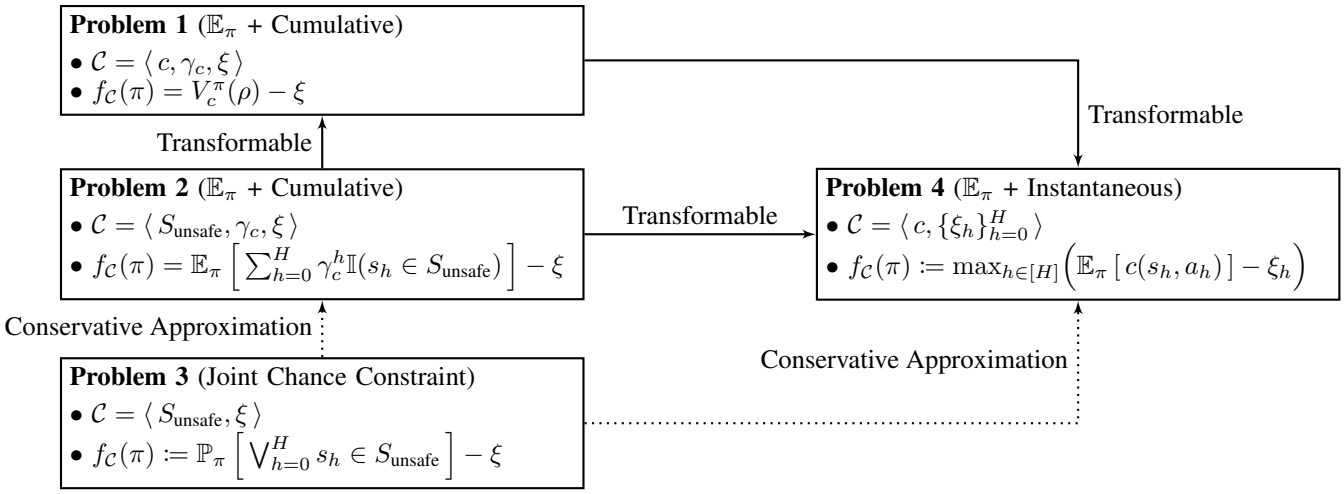

Figure 2: Relations among common safe RL formulations based on $\mathbb{E}_\pi$ and the one with chance constraints.

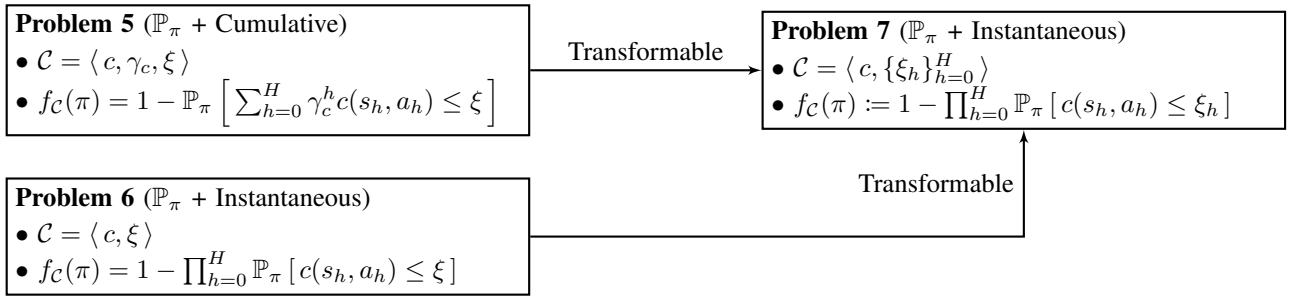

Figure 3: Relations among common safe RL formulations based on $\mathbb{P}_\pi$ (i.e., almost-surely constraints).

different classes of SCFs based on $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. For any SCF $f_{\mathcal{C}_1} \in \mathcal{F}_{\mathcal{C}_1}$, if there exists $f_{\mathcal{C}_2} \in \mathcal{F}_{\mathcal{C}_2}$ such that

$$\widehat{\Pi}(f_{\mathcal{C}_1}) = \widehat{\Pi}(f_{\mathcal{C}_2}),$$

then we say that the problem characterized by $\mathcal{M} \cup \mathcal{C}_1$ can be transformed into that characterized by $\mathcal{M} \cup \mathcal{C}_2$.

Definition 2 (Generalizability)**.** Let $N \in \mathbb{Z}_+$ denote a positive integer. Let $\{\mathcal{M} \cup \mathcal{C}_i\}_{i=1}^N$ denote a set of N CMDPs that are respectively characterized by different constraint tuples $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N$. Suppose that, for all $i \in [N]$, the problem characterized by $\mathcal{M} \cup \mathcal{C}_i$ can be transformed into that by $\mathcal{M} \cup \widetilde{\mathcal{C}}$. Then, we call the problem characterized by $\mathcal{M} \cup \widetilde{\mathcal{C}}$ is an identical or more general safe RL (IoMG-SafeRL) problem over the set of problems respectively characterized by $\mathcal{M} \cup \mathcal{C}_1, \mathcal{M} \cup \mathcal{C}_2, \ldots, \mathcal{M} \cup \mathcal{C}_N$.

Definition 3 (Conservative Approximation)**.** Let $\mathcal{M} \cup \mathcal{C}_1$ and $\mathcal{M} \cup \mathcal{C}_2$ be two CMDPs that are respectively characterized by different constraint tuples $\mathcal{C}_1$ and $\mathcal{C}_2$. Let $\mathcal{F}_{\mathcal{C}_1}$ and $\mathcal{F}_{\mathcal{C}_2}$ respectively denote two classes of SCFs based on $\mathcal{C}_1$ and $\mathcal{C}_2$. For any SCF $f_{\mathcal{C}_1} \in \mathcal{F}_{\mathcal{C}_1}$, if there exists $f_{\mathcal{C}_2} \in \mathcal{F}_{\mathcal{C}_2}$ such that

$$\widehat{\Pi}(f_{\mathcal{C}_1}) \supset \widehat{\Pi}(f_{\mathcal{C}_2}),$$

then we say that the problem characterized by $\mathcal{M} \cup \mathcal{C}_2$ is a conservative approximation of that characterized by $\mathcal{M} \cup \mathcal{C}_1$.

4.2 Preliminary Lemmas

To present the main theoretical results, we first list three necessary lemmas as preliminaries. The first lemma is based on Sootla *et al.* [2022b] and Wachi *et al.* [2023], which describes a theoretical connection between additive and instantaneous safety constraints.

Lemma 1. *Define a new variable η_h meaning the remaining safety budget associated with the discount factor γ_c such that*

$$\eta_h \coloneqq \gamma_c^{-h} \cdot \left(\xi - \sum_{h'=0}^{h-1} \gamma_c^{h'} c(s_{h'}, a_{h'}) \right), \ \forall h \in [H]. \quad (8)$$

Then, the following relation between additive and instantaneous constraints holds:

$$\sum_{h'=0}^{h-1} \gamma_c^{h'} c(s_{h'}, a_{h'}) \leq \xi \iff c(s_h, a_h) \leq \eta_h, \quad \forall h \in [H].$$

Proof. By definition, the new variable η_h satisfies the following recurrence formula:

$$\eta_{h+1} = \gamma_c^{-1} \cdot (\, \eta_h - c(s_h, a_h) \,) \quad \text{with} \quad \eta_0 = \xi. \quad (9)$$

Combining (9) and $c(s, a) \geq 0, \forall (s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\eta_{H+1} \geq 0 \iff \eta_h - c(s_h, a_h) \geq 0, \quad \forall h \in [H].$$

By definition of η_h, we have

$$\sum_{h'=0}^{H} \gamma_c^{h'} c(s_{h'}, a_{h'}) \leq \xi \iff \eta_{H+1} \geq 0. \tag{10}$$

Therefore, we obtain the desired lemma. □

Lemma 2. *Problem 1 can be transformed into Problem 4.*

Proof. By applying Lemma 1, the safety constraint of Problem 1 satisfies

$$V_c^\pi(\rho) \leq \xi \iff \mathbb{E}_\pi\left[\, c(s_h, a_h) \,\right] \leq \mathbb{E}_\pi\left[\, \eta_h \,\right],\ h \in [H].$$

Therefore, Problem 1 is identical to a special case of Problem 4 with $\xi_h \coloneqq \mathbb{E}_\pi[\eta_h]$ for all $h \in [H]$. Hence, we obtain the desired lemma. □

Lemma 3. *Problem 5 can be transformed into Problem 7.*

Proof. By applying Lemma 1, we have

$$\begin{aligned} &\mathbb{P}_\pi\left[\sum_{h=0}^{H} \gamma_c^h c(s_h, a_h) \leq \xi\right] = 1 \\ &\iff \mathbb{P}_\pi\left[\, c(s_h, a_h) \leq \eta_h \,\right] = 1, \quad \forall h \in [H]. \end{aligned}$$

Therefore, Problem 5 is identical to a special case of Problem 7 with $\xi_h \coloneqq \eta_h$ for all $h \in [H]$. Therefore, we obtain the desired lemma. □

4.3 The Two IoMG-SafeRL Problems

We now provide three theorems on the interrelations among the common safe RL problems. Crucially, we show that Problems 4 and 7 can be regarded as two IoMG-SafeRL problems of other problems. We also show that Problem 4 with $\gamma_c = 1$ is a conservative approximation of Problem 3. Conceptual illustrations are given in Figures 2 and 3.

Theorem 1. *Problem 4 is an IoMG-SafeRL problem over Problems 1 and 2.*

Proof. By Lemma 2, Problem 1 can be transformed into Problem 4. Also, Problem 2 can be easily transformed into Problem 1 by defining $c(s, a) \coloneqq \mathbb{I}(s \in S_{\text{unsafe}})$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. In summary, Problem 4 is an IoMG-SafeRL problem over Problems 1 and 2. □

Theorem 2. *Problem 4 with $\gamma_c = 1$ is a conservative approximation of Problem 3.*

Proof. This lemma mostly follows from Theorem 1 in Ono *et al.* [2015]. Regarding the constraint in Problem 3, we have the following chain of equations:

$$\begin{aligned} \mathbb{P}_\pi\left[\bigvee_{h=1}^{H} s_h \in S_{\text{unsafe}}\right] &\leq \sum_{h=1}^{H} \mathbb{P}_\pi\left[\, s_h \in S_{\text{unsafe}} \,\right] \\ &= \sum_{h=1}^{H} \mathbb{E}_\pi\left[\, \mathbb{I}(s_h \in S_{\text{unsafe}}) \,\right] \\ &= \mathbb{E}_\pi\left[\sum_{h=1}^{H} \mathbb{I}(s_h \in S_{\text{unsafe}})\right]. \end{aligned}$$

In the first step, we used Boole's inequality (i.e., $\Pr[A \cup B] \leq \Pr[A] + \Pr[B]$). The final term is the left-hand side of the constraint in Problem 2 with $\gamma_c = 1$, which implies that Problem 2 is a conservative approximation of Problem 3. Therefore, we obtained the desired theorem. □

Theorem 3. *Problem 7 is an IoMG-SafeRL problem over Problems 5 and 6.*

Proof. By Lemma 3, Problem 5 can be transformed into Problem 7. Also, Problem 6 is a special case of Problem 7 where ξ_h is a constant for all $h \in [H]$. Hence, Problem 7 is an IoMG-SafeRL problem over Problems 5 and 6. □

5 Discussion

We conclude by discussing the current state and future directions of safe RL based on constrained criteria.

5.1 Formulation and Algorithm Selection

As mentioned earlier, when we adopt the safe RL paradigm, there are two main steps: 1) *problem formulation* and 2) *policy optimization*. Based on the survey results, let us discuss how to solve safe RL problems via constraints.

Problem formulation. As presented in Section 3, constraint formulations in safe RL are divided into two classes: one based on $\mathbb{E}_\pi$ and the other based on $\mathbb{P}_\pi$. Constraint formulations with expectation $\mathbb{E}_\pi$ (i.e., Problems 1 - 4) represent the safety constraints using the expected value, which focuses on the "averaged" performance of safety. On the other hand, constraint formulations based on $\mathbb{P}_\pi$ (i.e., Problems 5 - 7) require an agent to guarantee safety almost surely. When the problem is formulated based on $\mathbb{P}_\pi$, the achieved safety level is usually higher by nature. Still, there is also a drawback that the reward performance is usually lower due to stricter safety constraints. Which formulation to adopt should depend on which level of safety is faced with the problem.

Policy optimization. When optimizing a policy, we must select a proper algorithm that corresponds to the problem formulation. Given the current situation, the easiest way is to use algorithms implemented in well-used OSS such as FSRL/OSRL [Liu *et al.*, 2023a] and SafePO/OmniSafe [Ji *et al.*, 2023]. Note that, as shown in Table 1, the algorithms implemented in the above OSS are mostly based on Problems 1 and 5. In the long run, however, the algorithms based on Problems 4 and 7 (i.e., IoMG-SafeRL problems) are also promising since instantaneous constraints are easier to handle than cumulative ones both theoretically and empirically.

When should your agent be safe? Another perspective in the algorithm selection should include *when* an agent needs to satisfy the safety constraint. Existing algorithms can be divided into two classes (see "safety" column in Table 1. The first class (w/o **T**) tries to achieve the required level of safety *after convergence* while encouraging safety during training. The algorithms in this class are usually based on Problems 1 or 5 where the safety constraints are represented using the additive safety cost structure. The second class (w/ **T**) tries to guarantee safety even *during training* as well as after convergence. The algorithms in this class are typically built upon

Problems 6 or 7 where the safety constraints are instantaneous. Which class of algorithm you should choose depends solely on the problem and application to be addressed.

5.2 Online RL vs. Offline RL

As shown in Table 1, most of the existing safe RL literature considers "online" RL settings where an agent interacts with the environment while learning its policy. A major advantage of safe "online" RL is that policies can be trained from scratch without data collected previously. On the other hand, "offline" RL [Levine *et al.*, 2020] is a framework to train a policy from a fixed amount of pre-collected data, potentially solving the fundamental issues regarding safety or risk. Offline RL is well-suited in the context of safe RL because the agent does not interact with the real environment during training; thus, the policy training does not essentially pose any risk. Hence, *safe offline RL* is a promising approach to achieve safety in RL if enough data or high-fidelity simulation is available. Although most of the existing safe offline RL literature is based on Problem 1 (e.g., Le *et al.* [2019], Lee *et al.* [2021]) as shown in Table 1, such research direction is expected to expand to other problem settings.

6 Conclusion

This paper provides a comprehensive survey of safe reinforcement learning based on constrained optimization criteria, with a particular focus on problem *formulations*. We present seven common constraint formulations and their associated representative algorithms in Section 3. Additionally, a curated selection of relevant literature is summarized according to the problem formulation in Table 1. In Section 4, we examine the theoretical relations among these formulations, offering readers a deeper understanding of safe RL. Furthermore, we describe the current status of safe RL research and outline potential future directions. Through this survey paper, we aim to foster a systematic understanding of constraint formulations and encourage further fundamental and applied research in safe reinforcement learning.

References

[Achiam *et al.*, 2017] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *ICML*, 2017.

[Afsar *et al.*, 2022] M Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

[Alshiekh *et al.*, 2018] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *AAAI*, 2018.

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

[Amani and Yang, 2022] Sanae Amani and Lin F Yang. Doubly pessimistic algorithms for strictly safe off-policy optimization. In *CISS*, 2022.

[Amani *et al.*, 2021] Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe reinforcement learning with linear function approximation. In *ICML*, 2021.

[Ames *et al.*, 2016] Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2016.

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

[As *et al.*, 2021] Yarden As, Ilnura Usmanova, Sebastian Curi, and Andreas Krause. Constrained policy optimization via Bayesian world models. In *ICLR*, 2021.

[Bai *et al.*, 2022] 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 *AAAI*, 2022.

[Berkenkamp *et al.*, 2017] Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *NeurIPS*, 2017.

[Bharadhwaj *et al.*, 2021] Homanga Bharadhwaj, Aviral Kumar, Nicholas Rhinehart, Sergey Levine, Florian Shkurti, and Animesh Garg. Conservative safety critics for exploration. In *ICLR*, 2021.

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

[Brunke *et al.*, 2021] 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.

[Bura *et al.*, 2022] Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning. In *NeurIPS*, 2022.

[Cheng *et al.*, 2019] 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 *AAAI*, 2019.

[Dai *et al.*, 2023] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. In *ICLR*, 2023.

[Dalal *et al.*, 2018] 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.

[Ding *et al.*, 2020] Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. In *NeurIPS*, 2020.

[Ding *et al.*, 2021] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *AISTAT*, 2021.

[Fisac *et al.*, 2018] 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.

[Fulton and Platzer, 2018] Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In *AAAI*, 2018.

[Garcıa and Fernández, 2015] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *JMLR*, 16(1):1437–1480, 2015.

[Geibel, 2006] Peter Geibel. Reinforcement learning for MDPs with constraints. In *ECML*, 2006.

[Gu *et al.*, 2022] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.

[Hambly *et al.*, 2023] Ben Hambly, Renyuan Xu, and Huining Yang. Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3):437–503, 2023.

[Hasanbeig *et al.*, 2020] Mohammadhosein Hasanbeig, Alessandro Abate, and Daniel Kroening. Cautious reinforcement learning with logical constraints. In *AAMAS*, 2020.

[Heger, 1994] Matthias Heger. Consideration of risk in reinforcement learning. In *ICML*, 1994.

[Ji *et al.*, 2023] Jiaming Ji, Jiayi Zhou, Borong Zhang, et al. OmniSafe: An infrastructure for accelerating safe reinforcement learning research. *arXiv preprint arXiv:2305.09304*, 2023.

[Jia *et al.*, 2020] Yan Jia, John Burden, Tom Lawton, et al. Safe reinforcement learning for sepsis treatment. In *IEEE ICHI*, 2020.

[Kim *et al.*, 2020] Youngmin Kim, Richard Allmendinger, and Manuel López-Ibáñez. Safe learning and optimization techniques: Towards a survey of the state of the art. In *TAILOR*, 2020.

[Kordabad *et al.*, 2022] Arash Bahari Kordabad, Rafael Wisniewski, and Sebastien Gros. Safe reinforcement learning using Wasserstein distributionally robust MPC and chance constraint. *IEEE Access*, 10:130058–130067, 2022.

[Le *et al.*, 2019] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *ICML*, 2019.

[Lee *et al.*, 2021] Jongmin Lee, Cosmin Paduraru, Daniel J Mankowitz, Nicolas Heess, Doina Precup, Kee-Eung Kim, and Arthur Guez. COptiDICE: Offline constrained reinforcement learning via stationary distribution correction estimation. In *ICLR*, 2021.

[Levine *et al.*, 2016] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *JMLR*, 17(1):1334–1373, 2016.

[Levine *et al.*, 2020] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[Li and Bastani, 2020] Shuo Li and Osbert Bastani. Robust model predictive shielding for safe reinforcement learning with stochastic dynamics. In *ICRA*, 2020.

[Li *et al.*, 2019] Yuanlong Li, Yonggang Wen, Dacheng Tao, and Kyle Guan. Transforming cooling optimization for green data center via deep reinforcement learning. *IEEE transactions on cybernetics*, 50(5):2002–2013, 2019.

[Liu *et al.*, 2020] Yongshuai Liu, Jiaxin Ding, and Xin Liu. IPO: Interior-point policy optimization under constraints. In *AAAI*, 2020.

[Liu *et al.*, 2021] Yongshuai Liu, Avishai Halev, and Xin Liu. Policy learning with constraints in model-free reinforcement learning: A survey. In *IJCAI*, 2021.

[Liu *et al.*, 2022] Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, and Ding Zhao. Constrained variational policy optimization for safe reinforcement learning. In *ICML*, 2022.

[Liu *et al.*, 2023a] Zuxin Liu, Zijian Guo, Haohong Lin, Yihang Yao, Jiacheng Zhu, Zhepeng Cen, Hanjiang Hu, Wenhao Yu, Tingnan Zhang, Jie Tan, et al. Datasets and benchmarks for offline safe reinforcement learning. *arXiv preprint arXiv:2306.09303*, 2023.

[Liu *et al.*, 2023b] Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. In *ICML*, 2023.

[Meng and Khushi, 2019] Terry Lingze Meng and Matloob Khushi. Reinforcement learning in financial markets. *Data*, 4(3):110, 2019.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[Moldovan and Abbeel, 2012] Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in markov decision processes. In *ICML*, 2012.

[Mowbray *et al.*, 2022] Max Mowbray, Panagiotis Petsagkourakis, Ehecatl Antonio del Rio-Chanona, and Dongda Zhang. Safe chance constrained reinforcement learning for batch process control. *Computers & chemical engineering*, 157:107630, 2022.

[Ono *et al.*, 2015] 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.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

[Pfrommer *et al.*, 2022] Samuel Pfrommer, Tanmay Gautam, Alec Zhou, and Somayeh Sojoudi. Safe reinforcement learning with chance-constrained model predictive control. In *L4DC*, 2022.

[Pham *et al.*, 2018] Tu-Hoa Pham, Giovanni De Magistris, and Ryuki Tachibana. Optlayer-practical constrained optimization for deep reinforcement learning in the real world. In *IEEE ICRA*, 2018.

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

[Shalev-Shwartz *et al.*, 2016] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

[Sharpe, 1966] William F Sharpe. Mutual fund performance. *The Journal of business*, 39(1):119–138, 1966.

[Shi *et al.*, 2023] Ming Shi, Yingbin Liang, and Ness Shroff. A near-optimal algorithm for safe reinforcement learning under instantaneous hard constraints. In *ICML*, 2023.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[Sootla *et al.*, 2022a] Aivar Sootla, Alexander Cowen-Rivers, Jun Wang, and Haitham Bou Ammar. Enhancing safe exploration using safety state augmentation. In *NeurIPS*, 2022.

[Sootla *et al.*, 2022b] Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H Mguni, Jun Wang, and Haitham Ammar. Sauté RL: Almost surely safe reinforcement learning using state augmentation. In *ICML*, 2022.

[Stooke *et al.*, 2020] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. In *ICML*, 2020.

[Sui *et al.*, 2015] Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with Gaussian processes. In *ICML*, 2015.

[Sutton and Barto, 1998] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[Tamar *et al.*, 2012] Aviv Tamar, Dotan Di Castro, and Shie Mannor. Policy gradients with variance related risk criteria. In *ICML*, 2012.

[Tessler *et al.*, 2019] Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *ICLR*, 2019.

[Thananjeyan *et al.*, 2021] Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. Recovery RL: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6(3):4915–4922, 2021.

[Thomas *et al.*, 2021] Garrett Thomas, Yuping Luo, and Tengyu Ma. Safe reinforcement learning by imagining the near future. In *NeurIPS*, 2021.

[Turchetta *et al.*, 2016] Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite Markov decision processes with Gaussian processes. In *NeurIPS*, 2016.

[Turchetta *et al.*, 2020] Matteo Turchetta, Andrey Kolobov, Shital Shah, Andreas Krause, and Alekh Agarwal. Safe reinforcement learning via curriculum induction. In *NeurIPS*, 2020.

[Wachi and Sui, 2020] Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained Markov decision processes. In *ICML*, 2020.

[Wachi *et al.*, 2018] Akifumi Wachi, Yanan Sui, Yisong Yue, and Masahiro Ono. Safe exploration and optimization of constrained MDPs using Gaussian processes. In *AAAI*, 2018.

[Wachi *et al.*, 2023] Akifumi Wachi, Wataru Hashimoto, Xun Shen, and Kazumune Hashimoto. Safe exploration in reinforcement learning: A generalized formulation and algorithms. In *NeurIPS*, 2023.

[Wang *et al.*, 2023] Yixuan Wang, Simon Sinong Zhan, Ruochen Jiao, Zhilu Wang, Wanxin Jin, Zhuoran Yang, Zhaoran Wang, Chao Huang, and Qi Zhu. Enforcing hard constraints with soft barriers: Safe reinforcement learning in unknown stochastic environments. In *ICML*, 2023.

[Wu *et al.*, 2021] Runzhe Wu, Yufeng Zhang, Zhuoran Yang, and Zhaoran Wang. Offline constrained multi-objective reinforcement learning via pessimistic dual value iteration. In *NeurIPS*, 2021.

[Wurman *et al.*, 2022] Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, et al. Outracing champion gran turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, 2022.

[Xu *et al.*, 2021] Tengyu Xu, Yingbin Liang, and Guanghui Lan. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *ICML*, 2021.

[Xu *et al.*, 2022] Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. Constraints penalized Q-learning for safe offline reinforcement learning. In *AAAI*, 2022.

[Yang *et al.*, 2020] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Projection-based constrained policy optimization. In *ICLR*, 2020.

[Yang *et al.*, 2021] Tsung-Yen Yang, Michael Y Hu, Yinlam Chow, Peter J Ramadge, and Karthik Narasimhan. Safe reinforcement learning with natural language constraints. In *NeurIPS*, 2021.

[Yu *et al.*, 2021] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[Yu *et al.*, 2022] Haonan Yu, Wei Xu, and Haichao Zhang. Towards safe reinforcement learning with a safety editor policy. In *NeurIPS*, 2022.

[Zhang *et al.*, 2020] Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. In *NeurIPS*, 2020.

[Zhang *et al.*, 2022] Linrui Zhang, Li Shen, Long Yang, Shixiang Chen, Bo Yuan, Xueqian Wang, and Dacheng Tao. Penalized proximal policy optimization for safe reinforcement learning. In *IJCAI*, 2022.

[Zhao *et al.*, 2023a] Weiye Zhao, Rui Chen, Yifan Sun, Tianhao Wei, and Changliu Liu. State-wise constrained policy optimization. *arXiv preprint arXiv:2306.12594*, 2023.

[Zhao *et al.*, 2023b] Weiye Zhao, Tairan He, Rui Chen, Tianhao Wei, and Changliu Liu. State-wise safe reinforcement learning: A survey. In *IJCAI*, 2023.