POCE: Primal Policy Optimization with Conservative Estimation for Multi-constraint Offline Reinforcement Learning

Jiayi Guan[1†] , Li Shen[2†], Ao Zhou[1], Lusong Li[2],Han Hu[3],
Xiaodong He[2], Guang Chen[1‡], Changjun Jiang[1]
[1]Tongji University, [2] JD Explore Academy, [3] Beijing Institute of Technology

Abstract

*Multi-constraint offline reinforcement learning (RL) promises to learn policies that satisfy both cumulative and state-wise costs from offline datasets. This arrangement provides an effective approach for the widespread application of RL in high-risk scenarios where both cumulative and state-wise costs need to be considered simultaneously. However, previously constrained offline RL algorithms are primarily designed to handle single-constraint problems related to cumulative cost, which faces challenges when addressing multi-constraint tasks that involve both cumulative and state-wise costs. In this work, we propose a novel **P**rimal policy **O**ptimization with **C**onservative **E**stimation algorithm (POCE) to address the problem of multi-constraint offline RL. Concretely, we reframe the objective of multi-constraint offline RL by introducing the concept of Maximum Markov Decision Processes (MMDP). Subsequently, we present a primal policy optimization algorithm to confront the multi-constraint problems, which improves the stability and convergence speed of model training. Furthermore, we propose a conditional Bellman operator to estimate cumulative and state-wise Q-values, reducing the extrapolation error caused by out-of-distribution (OOD) actions. Finally, extensive experiments demonstrate that the POCE algorithm achieves competitive performance across multiple experimental tasks, particularly outperforming baseline algorithms in terms of safety. Our code is available at github.POCE.*

1. Introduction

Reinforcement learning (RL) have achieved remarkable achievements in the domains of policy games [4, 19, 42], robotics [2, 17, 39, 50], and recommendation systems [5, 7]. However, safety concerns remain a primary challenge to the real-world deployment of RL, particularly in scenarios with high safety requirements [15, 21, 38, 53]. These scenarios demand attention to both cumulative and state-wise safety constraints. For example, in the context of autonomous driving, the objective is not only to mitigate instances of prolonged hazardous driving but also to prevent immediate collisions [16, 33, 46, 58]. Multi-constraint offline RL is a promising and potentially effective approach to address the aforementioned issues. It learns policies that satisfy both cumulative and state-wise safety constraints from pre-collected offline datasets.

Currently, several works leverage the framework of Constrained Markov Decision Processes (CMDP) to model and address the problem of learning cost-constrained strategies under offline dataset, achieving promising results in handling tasks with cumulative cost constraints [22, 30, 32, 44]. However, these approaches face challenges when dealing with problems that involve constraints on both cumulative and state-wise cost simultaneously [6, 13, 30]. Additionally, for the aforementioned multi-constraint problems, methods such as the Lagrange multiplier and penalty function method are commonly adopted [12, 24, 31, 36]. These methods are sensitive to initial values, which leads to unstable model training. Moreover, they introduce additional penalty coefficients or Lagrange multipliers, increasing the cost of model hyperparameter tuning. On the other hand, the extrapolation error caused by out-of-distribution (OOD) actions in offline settings also significantly affects the performance of optimization algorithms [14, 29, 54]. Therefore, this work focuses on solving multi-constraint optimization problems involving cumulative and state-wise costs in the offline setting.

To solve the aforementioned multi-constraint optimization problem in the offline setting, we propose a novel **P**rimal policy **O**ptimization with **C**onservative **E**stimation algorithm (POCE). Concretely, we introduce the Maximum Markov Decision Process to represent the state-wise cost. Based on this, we redefine a novel multi-constrained offline RL objective for tasks that involve both cumulative and state-wise costs. Subsequently, we propose a pri-

† Equal Contribution;
‡ Corresponding author: guangchen@tongji.edu.cn.

mal policy optimization approach to address the multi-constrained offline RL task. Additionally, we present a conditional Bellman operator to mitigate the extrapolation error of the cumulative and state-wise cost Q-values caused by OOD actions. Finally, extensive experiments demonstrate that the POCE algorithm achieves competitive performance across multiple experimental tasks, particularly outperforming baseline algorithms in terms of safety.

The main contributions of this work are listed as follows:

- To the best of my knowledge, we are the first to formulate a multi-constrained offline RL objective to address the policy optimization problem involving cumulative and state-wise costs under offline settings.
- We introduce a primal policy optimization method to address the multi-constrained optimization problem, which improves the stability of the algorithm and reduces the tuning cost associated with the introduction of additional hyperparameters.
- We propose a novel conditional Bellman operator to address the Q-value iteration problem for cumulative and state-wise costs, This operator ensures that the cost Q-values are not underestimated while guaranteeing convergence to a unique fixed point.
- Extensive comparisons and ablation experiments demonstrate that the POCE algorithm delivers competitive performance, particularly in terms of safety.

2. Related Work

In this section, we extensively discuss the related work on multi-constrained offline RL. We primarily focus on two aspects: constrained RL and constrained offline RL.

Constrained RL is an approach to address constrained policy optimization problems by introducing costs corresponding to rewards based on standard RL. Currently, there are numerous constrained RL algorithms based on the CMDP framework for handling cumulative costs [11, 38, 48]. Among them, the algorithms based on the primal-dual methods have been widely applied to solve constrained RL problems, such as PDO [9], RCPO [37], and CVPO [25]. Furthermore, there also have been successful attempts in addressing constrained RL tasks using second-order approximation [1, 49] and two-step projection methods [47, 59], resulting in excellent performance. Additionally, some works have combined commonly used constrained methods such as interior-point [24], conditional value-at-risk [8, 34, 52], and penalty function methods [55] to solve constrained RL tasks. On the other hand, some studies focus on the state-wise cost [60, 61], such as USL [56], which achieves state-wise cost constraints through hierarchical implementation. Additionally, some works model state-wise cost as Gaussian processes and satisfy state-wise constraints through planning with nearby states [40, 41].

Constrained Offline RL is to solve the problem of policy optimization with cost constraints under an offline dataset. After the proposal of a constrained offline RL algorithm with constrained penalties [44], several algorithms have been developed to achieve cost-constrained policy learning from offline datasets using different methods. Some works combine stationary distributions and linear programming to solve for the optimal stationary distribution, thereby enabling the learning of policies that satisfy cumulative cost constraints from offline dataset [22, 32]. Additionally, a method based on variational inference and pessimistic conservative estimation has been used to address constrained offline RL tasks [13]. On the other hand, some works have addressed the optimization problem of satisfying cumulative costs under offline data from the perspective of sequential decision-making using the Transformer model [27, 51, 57].

In summary, our work is distinct from existing studies in two main aspects, which also represent the core challenges we meet. Firstly, our work addresses the multi-constrained optimization problem of simultaneous handling of cumulative and state-wise costs in high-risk scenarios. Secondly, our work deals with the multi-constrained optimization problem in the offline setting. To the best of my knowledge, this is the first attempt to simultaneously handle cumulative state-wise costs in the offline setting.

3. Preliminaries

In this section, we present the fundamental concepts and background of constrained RL and constrained offline RL. Subsequently, we rethink the existing works and highlight the purpose and significance of our work.

3.1. Constrained RL and Constrained Offline RL

CMDP provides a theoretical framework to solve constrained RL problems [3]. It is defined as a tuple $(\mathcal{S}, \mathcal{A}, C, P, r, \rho_0, \gamma)$, where $\mathcal{S} \in \mathbb{R}^n$ is the state space, $\mathcal{A} \in \mathbb{R}^m$ is the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition kernel, which specifies the transition probability $p(s_{t+1}|s_t, a_t)$ from state s_t to state s_{t+1} under the action a_t, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represents the reward function, C is the set of costs $\{c_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+, i = 1, 2, \cdots, m\}$ for violating m constraints, $\gamma \in [0, 1)$ is the discount factor, and $\rho_0 : \mathcal{S} \to [0, 1]$ is the distribution of initial states. The policy π is a probability distribution mapping the state s_t to the action a_t. We use shorthand $r_t = r(s_t, a_t)$ and $c_{i,t} = c_i(s_t, a_t)$ for simplicity. The common objective of constrained RL is to maximize the cumulative reward while satisfying the cumulative cost constraint.

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right], \text{s.t. } \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t c_{i,t}\right] \leq \bar{c}_i, \quad (1)$$

where the $\tau = \{s_0, a_0, \cdots\} \sim \pi$ denotes the trajectory. $\bar{c}_i$ is the cost threshold of the i-th cumulative cost constraint.

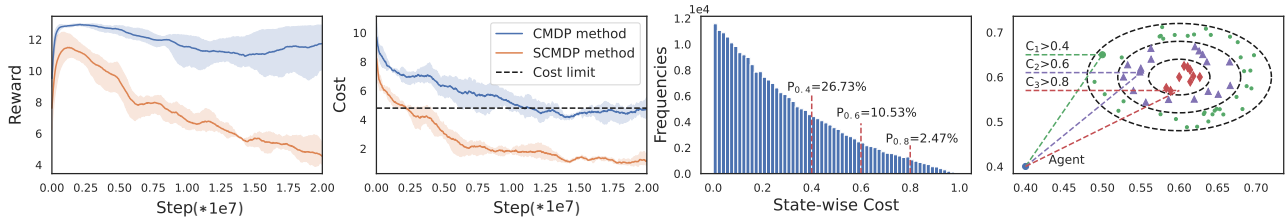

Figure 1. The left two sub-figures illustrate the reward and cost curves under the cumulative cost constraint and state-wise cost constraint method for the *PointGoal* task. The third sub-figure illustrates the state-wise cost violation rate under different state-wise cost thresholds when satisfying the cumulative cost constraint using the cumulative cost constraint algorithm. The last subfigure depicts the region range of different state-wise costs. The threshold for cumulative cost is set as $\bar{c}_i = 5$.

On the other hand, to ensure that the agent satisfies cost constraints at each time step during the execution of the policy, a State-wise Constrained Markov Decision Process (SCMDP) is proposed [61]. Unlike the cumulative cost constraint in CMDP mentioned earlier, SCMDP requires the agent to satisfy the cost constraint in every state. The objective of SCMDP be formulated as follows:

$$\begin{aligned} \pi^* &= \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right], \\ \text{s.t. } \forall t \geq 0,\ & \mathbb{E}_{s_t\sim p(\cdot|s_t,a_t),a_t\sim\pi(\cdot|s_t)}(c_{i,t}) \leq \bar{D}_i. \end{aligned} \tag{2}$$

where t represents the t-th step, and $\bar{D}_i$ denotes the cost threshold of the i-th sate-wise cost constraint.

Although the aforementioned online-constrained RL methods can ensure the safety of the agent during the testing and deployment phases, their safety during the training process faces challenges. Constrained offline RL is a method of learning a policy that satisfies cost constraints from offline datasets without interacting with the environment [22, 32]. This paradigm not only addresses the safety concerns during the testing and deployment phases but also ensures the safety of the training process.

3.2. Rethinking to Constrained RL

We rethink the application requirements of RL algorithms in real-world scenarios. In the field of autonomous driving, we allow intelligent agent vehicles to exceed the speed limit or change lanes within a limited timeframe, but we do not permit long-term speeding or crossing lane lines, nor do we allow any collisions. Similarly, in the domain of quantitative investing, it is necessary to consider not only the overall investment limit but also restrictions on individual investment amounts. These cases require us to simultaneously consider cumulative and state-wise costs.

We conduct experiments on the constraint RL algorithms of CMDP and SCMDP in the typical experimental scenario $PointGoal$. The experimental results are shown in Fig. 1. The results shown in Fig. 1 indicate that in the constraint RL for handling cumulative costs, the probability of the agent's state-wise cost exceeding 0.6 is more than 10% when the cumulative cost constraint is satisfied. This suggests that although the constraint RL algorithm for addressing cumulative costs can ensure that the agent satisfies the cumulative cost constraint, it is hard to guarantee that the state-wise cost remains within a controllable range. Additionally, we can also observe from the results in Fig. 1 that the constraint RL algorithm for handling state-wise costs does not ensure that all states satisfy the cost constraint, and it leads to a significant loss of reward.

In summary, the current individual treatment of cumulative and state-wise costs in constraint RL does not adequately meet the requirements of application scenarios that involve both cumulative and state-wise costs. Inspired by CMDP, SCMDP, and offline RL, we propose a novel multi-constraint offline RL task for addressing the simultaneous treatment of cumulative and state-wise costs in RL application scenarios.

4. Methodology

In this section, we provide a detailed exposition of primal policy optimization with conservative estimation algorithm for multi-constraint offline RL. Firstly, we define the multi-constraint offline RL task and reframe the objective of multi-constraint offline RL by introducing the concept of maximizing the MDP (MMDP). Based on this, we propose a primal policy optimization with conservative estimation algorithm for multi-constraint offline RL. Additionally, we propose a conditional Bellman operator to handle the extrapolation error in cost Q-values caused by OOD actions.

Definition 1 *In the offline setting, the multi-constraint safe RL objective that encompasses cumulative and state-wise cost constraints is as follows:*

$$\begin{aligned} \pi^* &= \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right], \\ \text{s.t.}\quad & \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t c_i(s_t,a_t)\right] \leq \bar{c}_i, \\ \forall t \geq 0,\ & \mathbb{E}_{s_t\sim p(\cdot|s_t,a_t),a_t\sim\pi(\cdot|s_t)}[c_i(s_t,a_t)] \leq \bar{D}_i. \end{aligned} \tag{3}$$

4.1. Multi-constraint Offline Safe RL

Based on the analysis and discussion in Section 3.2, it is known that existing constrained RL algorithms meet chal-

lenges when dealing with tasks that involve both cumulative and stat-wise cost constraints. Therefore, we define a multi-constraint offline RL to address the problem of simultaneously incorporating cumulative and state-wise costs in the offline setting.

We find that directly addressing the state-wise cost constraint as shown in Eq. (3) is extremely challenging. Inspired by SCPO [61], we transform the state-wise cost into a state cost increment. Concretely, we extend the objective of multi-constraint offline RL defined in Definition 1 by introducing a set of maximum state-wise costs $M_{i,t}$ and state-wise cost increments $D_{i,t}$. Additionally, we obtain an augmented state $\hat{s}_t = (s_t, M_{i,t})$ by supplementing the state s_t with the maximum state cost $M_{i,t}$. The maximum state-wise costs and state-wise cost increments are expressed as:

$$D_i(\hat{s}_t, a_t) = \max\{c_i(s_t, a_t) - M_{i,t}, 0\}, \tag{4}$$

$$M_{i,t} = \sum_{k=0}^{k=t-1} D_i(\hat{s}_k, a_k), \tag{5}$$

where $M_{i,t} = M_i(\hat{s}_t, a_t)$ represents the maximum state-wise cost of the i-th cost at step t, and $D_{i,t} = D_i(\hat{s}_t, a_t)$ represents the increment of the state-wise of the i-th cost at step t. The initial values of the maximum state-wise cost and the increment of the state-wise cost are defined as $M_{i,0} = 0$ and $D_i(\hat{s}_0, a_0) = c_i(s_0, a_0)$. Combining the maximum state-wise cost and the increment of the state-wise cost as defined in Eq. (4) and (5), we derive Lemma 4.1 based on Definition 1. Due to the space limitation, proofs and discussions are in **Appendix A.1**.

Lemma 4.1 *The objective of the multi-constraint offline RL with cumulative and state-wise costs can be formulated as:*

$$\begin{aligned} &\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(\hat{s}_t, a_t)\right], \\ &\text{s.t.}\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t c_i(\hat{s}_t, a_t)\right] \le \bar{c}_i,\ \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} D_i(\hat{s}_t, a_t)\right] \le \bar{D}_i. \end{aligned} \tag{6}$$

where $r(\hat{s}_t, a_t) \triangleq r(s_t, a_t)$ *and* $c_i(\hat{s}_t, a_t) \triangleq c_i(s_t, a_t)$.

We define the Q-value for reward on policy π as $Q^r(\hat{s}, a) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t r(\hat{s}_t, a_t)|\hat{s}_0 = \hat{s}, a_0 = a]$, and the Q-value of cost is defined similarly. Additionally, we define the Q-vlaue for cost increments on policy π as $Q^{D_i}(\hat{s}, a) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} D_i(\hat{s}_t, a_t)|\hat{s}_0 = \hat{s}, a_0 = a]$. Through theoretical derivation, we discovered an inherent connection between the state-wise and cumulative cost constraint, as mentioned in Remark 4.2. Proofs and discussions are in **Appendix A.2**.

Remark 4.2 *When the state-wise cost satisfies the cost constraint* $\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} D_i(\hat{s}_t, a_t)\right] \le \bar{D}_i$ *and the cost threshold of state-wise cost satisfies the condition* $\bar{D}_i \le (1-\gamma)\bar{c}_i$*, then the cumulative cost also satisfies the cost constraint* $\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t c_i(\hat{s}_t, a_t)\right] \le \bar{c}_i$.

4.2. Primal Policy Optimization

Due to the difficulty in directly handling the multi-constrained maximization reward objective stated in Lemma 4.1 under continuous state and action spaces, we are inspired by DDPG [35] and rewrite Eq. (6) as:

$$\begin{aligned} &\pi^* = \arg\max_{\pi} \mathbb{E}_{\hat{s}\sim\mathcal{D}, a\sim\pi}\left[Q^r_{\phi_r}(\hat{s}, a)\right], \\ &\text{s.t. } \mathbb{E}_{\substack{\hat{s}\sim\mathcal{D}\\ a\sim\pi}}\left[Q^{c_i}_{\phi_{c_i}}(\hat{s}, a)\right] \le \bar{c}_i,\ \mathbb{E}_{\substack{\hat{s}\sim\mathcal{D}\\ a\sim\pi}}\left[Q^{D_i}_{\phi_{D_i}}(\hat{s}, a)\right] \le \bar{D}_i, \end{aligned} \tag{7}$$

where $Q^r_{\phi_r}$, $Q^{c_i}_{\phi_{c_i}}$, and $Q^{D_i}_{\phi_{D_i}}$ represent the parameterized reward Q-value, cumulative and state-wise cost Q-value. For the multi-constrained optimization as shown in Eq. 7, the common practice is to apply the Lagrange multiplier method to transform the constrained problem into an unconstrained one [23, 37]. However, this conventional primal-dual method is sensitive to initial values and learning rates, resulting in significant costs during hyperparameter tuning. Inspired by CRPO [45], we directly optimize the aforementioned multi-constrained problem using the policy gradient of the primal problem based on stochastic approximation theory. This method avoids the training instability and tuning cost caused by additional Lagrange multipliers. On the other hand, OOD actions in offline settings have a significant impact on the accuracy of Q-value estimation. Therefore, to mitigate the extrapolation errors caused by OOD actions, we employ a conservative estimation method to estimate the Q-values of reward, cumulative cost and state-wise cost separately.

Based on the above analysis, we propose a **P**rimal policy **O**ptimization with **C**onservative **E**stimation (POCE) for multi-constraints offline RL tasks. We divide the POCE algorithm into two steps: Q-value estimation and policy update. In the Q-value estimation step, we present a conservative estimation method to estimate the Q-values of reward, cumulative cost, and state-wise cost separately. The specific estimation methods are detailed in Section 4.3. In the policy update step, we evaluate whether the current policy satisfies the cost constraints based on the Q-values of cumulative and state-wise cost. Subsequently, we update the policy by maximizing the reward or minimizing the cost. Concretely, we evaluate whether the cumulative cost $\mathbb{E}_{\hat{s}\sim\mathcal{D}, a\sim\pi}\left[Q^{c_i}_{\phi_{c_i}}(\hat{s}, a)\right] \le \bar{c}_i$ and state-wise cost $\mathbb{E}_{\hat{s}\sim\mathcal{D}, a\sim\pi}\left[Q^{D_i}_{\phi_{D_i}}(\hat{s}, a)\right] \le \bar{D}_i$ satisfy the constraints. If both constraints are satisfied, the constrained problem described in Eq. 7 is transformed into an unconstrained problem of maximizing the reward. If either constraint is not satisfied, the constrained problem in Eq. 7 is transformed into minimizing either the cumulative or state-wise cost. Therefore, the objective of policy updating can be expressed as:

$$\mathcal{L}(\theta) = \begin{cases} \arg\max_{\theta} \mathbb{E}\left[Q^r_{\phi_r}(\hat{s}, \pi_\theta(a|\hat{s}))\right], (\pi_\theta \in \pi_s) \\ \arg\min_{\theta} \mathbb{E}\left[Q^{\chi}_{\phi_\chi}(\hat{s}, \pi_\theta(a|\hat{s}))\right], (\pi_\theta \notin \pi_s) \end{cases}, \tag{8}$$

where π_s represents the safety policy, $\pi_\theta \in \pi_s$ indicates that both the cumulative and state-wise cost Q-values under the current policy π_θ meet the constraints, and $\pi_\theta \notin \pi_s$ indicates that at least one cost Q-value under the policy π_θ does not satisfy the constraints. The χ is either c_i or D_i. Note that if both constraints are not satisfied, we choose either one and minimize its cost.

4.3. Conservatively Estimate for Q-values

Due to the fact that the POCE algorithm evaluates whether the cost meets the constraints and computes the policy gradients are both based on the Q-values, the accuracy of Q-value estimation directly affects the algorithm's performance. Therefore, in the Q-value estimation step, we are committed to providing accurate Q-values. Considering that the experience Bellman iteration always selects action-state pairs with the maximum Q-values, the erroneous estimation of Q-values for OOD actions is difficult to correct in the offline setting, which leads to the overestimation of reward Q-values. To address the issue of reward Q-value overestimation caused by OOD actions, we follow the pessimistic estimation approach of CQL [20]. We incorporate a penalty term for the marginal distribution of unseen action based on the experience Bellman iteration and maximize the Q-value of the current policy to achieve a conservative estimation of the reward Q-values. Then, the loss function for the reward Q-value iteration can be expressed as follows:

$$\begin{aligned}\mathcal{L}(\phi_r) = \arg\min \frac{1}{2}\mathbb{E}_{\hat{s},a\sim D}\left[\left(Q^r_{\phi_r}(\hat{s},a) - \hat{\mathcal{T}}_{\mathcal{B}}\hat{Q}^r(\hat{s},a)\right)^2\right] \\ + \kappa\left[\mathbb{E}_{\substack{\hat{s}\sim\mathcal{D}\\ a\sim\pi_{\mathcal{M}}(\cdot|\hat{s})}} Q^r_{\phi_r}(\hat{s},a) - \mathbb{E}_{\substack{\hat{s}\sim\mathcal{D}\\ a\sim\hat{\pi}_\theta(\cdot|s)}} Q^r_{\phi_r}(\hat{s},a)\right],\end{aligned} \tag{9}$$

where κ is a tradeoff factor. Eq. (9) utilizes the empirical Bellman operator $\hat{\mathcal{T}}_{\mathcal{B}}$ instead of the actual Bellman Operator $\mathcal{T}_{\mathcal{B}}$. $\pi_{\mathcal{M}}$ is the marginal distribution of unseen actions.

On the other hand, to prevent unsafe actions during policy execution, we need to avoid underestimating the cost Q-values of action-state pairs while ensuring that the estimated cost Q-values do not significantly deviate from the true Q-values. Obviously, our estimation method for reward Q-values is no longer applicable to the estimation of Q-values for cumulative and state-wise costs. Inspired by MCQ [28], we propose a conditional Bellman operator for iterative estimation of cost Q-values. Concretely, when an action belongs to the in-distribution actions, we still employ the Bellman operator for Q-value iteration. However, when an action belongs to the OOD actions, we provide a pseudo target for the iterated Q-values. Then, the conditional Bellman iteration equation can be written as:

$$\mathcal{T}_{\mathcal{CB}}Q(\hat{s},a) = \begin{cases}\chi(\hat{s},a) + \gamma\mathbb{E}_{\hat{s}'}\max\limits_{a'} Q(\hat{s}',a'), & (a\in\pi_\beta(a|\hat{s}))\\ \max\limits_{\hat{a}\sim\pi_\beta} Q(\hat{s},\pi_\beta(a|\hat{s})), & (a\notin\pi_\beta(a|\hat{s}))\end{cases}, \tag{10}$$

Algorithm 1: POCE

Input: Dataset $\mathcal{D} = \{(\hat{s},a,r,c_i,D_i,\hat{s}',d)_i\}_{i=0}^n$
Output: Policy network parameters θ;
Parameters for the reward Q-values ϕ_r, ϕ_r^t, ;
Parameters for the cumulative cost Q-values ϕ_{c_i}, $\phi_{c_i}^t$;
Parameters for the incremental cost Q-values ϕ_{D_i}, $\phi_{D_i}^t$;
Parameters of the behavioral policy network ω;
for *each batch* **do**
  Sample a batch of transitions $(\hat{s},a,r,\hat{s}',c_i,D_i,d)$ from the buffer $\mathcal{D}$;
  `// Q-value Estimation Step:`
  Update the behavioral policy ω via the Eq. (12);
  Update the reward Q-values ϕ_r via the Eq. (9);
  Update the cumulative and incremental cost Q-values ϕ_{c_i} and ϕ_{D_i} via the Eq. (11);
  `// Policy Update Step:`
  Compute the cost constraint:
  $\bar{Q}^{c_i}(s,a) = \max_{(s,a)\sim D} Q^{c_i}_{\phi_{c_i}}(s,a)$,
  $\bar{Q}^{D_i}(s,a) = \max_{(s,a)\sim D} Q^{D_i}_{\phi_{D_i}}(s,a)$;
  if $\bar{Q}^{c_i}(s,a) \le \bar{c}_i$ *and* $\bar{Q}^{D_i}(s,a) \le \bar{D}_i$ **then**
    Take policy updates toward maximize reward
    $\theta \leftarrow \arg\max_\theta \mathbb{E}\left[Q^r_{\phi_r}(\hat{s},\pi_\theta(a|\hat{s}))\right]$;
  else
    Take policy updates to minimize cost
    $\theta \leftarrow \arg\min_\theta \mathbb{E}\left[Q^\chi_{\phi_\chi}(\hat{s},\pi_\theta(a|\hat{s}))\right]$;
  $\phi_r^t \leftarrow \tau\phi_r + (1-\tau)\phi_r^t$;
  $\phi_{c_i}^t \leftarrow \tau\phi_{c_i} + (1-\tau)\phi_{c_i}^t$;
  $\phi_{D_i}^t \leftarrow \tau\phi_{D_i} + (1-\tau)\phi_{D_i}^t$;

where π_β is the behavioral policy of the sample data. The χ represents the cost c_i or cost increment D_i. Note that an action is considered an in-distribution action if it belongs to the behavioral policy of the sample data; otherwise, it is considered an OOD action.

Proposition 4.3 *Within the scope of behavioral policies π_β, the conditional Bellman operator $\mathcal{T}_{\mathcal{CB}}$ is a γ-contractive operator under the $\mathcal{L}_{+\infty}$ norm, and any initial Q-value can converge to a unique fixed point through $\mathcal{T}_{\mathcal{CB}}$.*

Proposition 4.4 *As the unique fixed point of the conditional Bellman operator, $Q_{\mathcal{T}_{CB}}$ is bounded within the range of behavioral policies π_β, with $Q_{\mathcal{T}_{CB}} \in [Q_{\pi_\beta}, Q_{\pi^*_\beta}]$. Here, Q_{π_β} is the Q-value of the behavioral policy, and $Q_{\pi^*_\beta}$ is the Q-value of the optimal policy.*

Proposition 4.3 demonstrates that the Q-values converge to a unique fixed point through conditional Bellman iteration. Proposition 4.4 indicates that within the scope of supported behavioral policies, the Q-values obtained through conditional Bellman iteration does not exhibit substantial underestimation nor significant deviation from the true values. The proofs and discussions of Proposition 4.3 and 4.4 are

in **Appendix A.3** and **Appendix A.4.** Based on the conditional Bellman operator, we constructed an iterative loss function for the Q-values of cumulative costs. Concretely, we sampled OOD actions for state s using the current policy and assigned pseudo targets to the OOD actions under state s based on the current Q-value network.

$$\mathcal{L}(\phi_\chi) = \arg\min \lambda \mathbb{E}_{\hat{s},a\sim\mathcal{D}}\left[\left(Q^\chi_{\phi_\chi}(\hat{s},a) - \hat{\mathcal{T}}_\mathcal{B}\hat{Q}^\chi(\hat{s},a)\right)^2\right] + (1-\lambda)\mathbb{E}_{\hat{s}\sim\mathcal{D},a^o\sim\pi}\left[\left(Q^\chi_{\phi_\chi}(\hat{s},a^o) - y\right)^2\right], \tag{11}$$

where λ is a hyperparameter that balances the in-distribution and OOD actions. The χ is either c_i or D_i The Pseudo targets for OOD actions $y = \mathbb{E}_{\{\hat{a}_i\}^N\sim\pi_\beta}[\max_{\hat{a}\sim\{\hat{a}_i\}^N} Q(\hat{s},\hat{a})]$.

4.4. Practical Algorithm

To facilitate the understanding of the implementation process of the POCE algorithm, we provide a detailed explanation of a practical instance of the POCE algorithm. The pseudo-code for the POCE algorithm is shown in Algorithm 1. In the Q-value estimation step, we employ conservative estimation methods to update the Q-values of reward, cumulative cost, and state-wise cost, respectively. Additionally, since the conditional Bellman operator requires the behavioral policy π_β of the sample data, which is unknown in the offline setting, we utilize the learned behavioral policy $\hat{\pi}_\beta$ from a conditional variational autoencoder (CVAE) as a substitute for the behavioral policy π_β. Then, the loss function for the CVAE can be written as:

$$\mathcal{L}(\omega) = \arg\min \mathbb{E}_{(\hat{s},a)\sim\mathcal{D},z\sim E_\omega(\hat{s},a)}\left[(a - D_\omega(\hat{s},z))^2 + D_{KL}(z||\mathcal{N}(0,I))\right]. \tag{12}$$

In the policy update step, we evaluate whether the cost constraints are satisfied and update the policy network. Due to the limited number of action-sate pairs we collect, it is challenging to ensure that we sample action-state pairs with high accumulated or state-wise costs on every trajectory during the evaluation of cost constraints. Therefore, we scale the evaluation condition using the maximum Q-value when assessing whether the cost constraints are met.

5. Experimental Evaluation

In this section, we conduct comprehensive comparative experiments between POCE and previous offline safe RL methods using datasets from a range of domains encompassing diverse action spaces and observation dimensions.

5.1. Task and Baseline

Task and Dataset. To assess the performance of POCE in various tasks across different domains, we selected three widely adopted tasks [18, 26] from the Safety-gym and Mujoco domains as experimental tasks in this work. Concretely, we chose the *PointGoal* and *CarGoal* tasks in the safety-gym domain. These tasks require the agent to navigate the *Point* or *Car* through interference and avoid hazardous obstacles to reach a target location. Additionally, we select the *AntVelocity* task in the Mujoco domain which requires the *Ant* agent to walk or run within a specified velocity range.

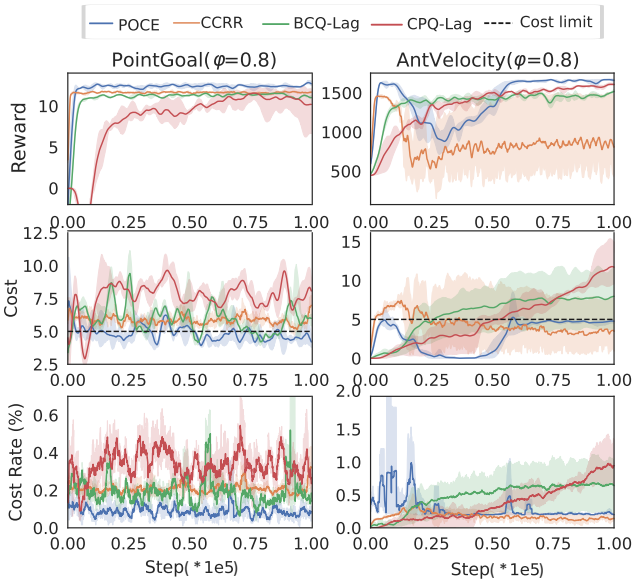

Figure 2. The figure depicts the reward and cumulative cost curves and the curve representing the violation rate of state costs. The shaded areas on the curves represent the variance obtained from online testing conducted with three random seeds. The safety factor for this experiment is $\varphi = 0.8$, and the cumulative cost threshold is set at $\bar{c}_i = 5$. In the *PointGoal* task, the state-wise cost threshold is set to $\bar{D}_i = 0.8$, while in the *AntVelocity* task, the state-wise cost threshold is $\bar{D}_i = 0.2$.

To the best of my knowledge, there is currently no standardized dataset for multi-constraint offline RL. To facilitate further research and reproducibility of this work, we revise the standard experimental scenarios of constrained RL by adjusting the cost values in the environment to conform to the range of $c_i \in [0,1]$ rather than $c_i \in \{0,1\}$. Moreover, drawing inspiration from work [13], we introduce a safety factor φ to differentiate the sample data of different behaviors. Subsequently, we collect diverse sample data based on different behaviors. The cost allocation for each task and the methods of diverse data collection are in **Appendix B.1**.

Baselines. Due to the lack of constrained offline RL algorithms that can simultaneously handle both cumulative and state-wise cost constraints. Therefore, we enhance existing constrained offline RL methods to develop multi-constraint offline RL algorithms as the experimental baselines. BCQ-Lag is an improved constrained offline RL algorithm that combines the Lagrange multiplier method with the BCQ [10] algorithm. We introduce two Lagrange multipliers to handle cumulative and state-wise costs re-

Table 1. The performance of the baseline and POCE algorithms is evaluated across samples with varying safety factors. The results from experiments involving 20 episodes are conducted with 3 random seeds. The data with a shaded background indicated the satisfaction of cumulative cost constraints, while the data highlighted in bold represents the highest reward return obtained under the satisfaction of cumulative cost constraints. The state-wise cost threshold for this experiment is $\bar{D}_i = 0.8$, and the cumulative cost threshold is $\bar{c}_i = 5$.

Method	Metrics	PointGoal			CarGoal			Mean
		φ=0.8	φ=0.6	φ=0.4	φ=0.8	φ=0.6	φ=0.4	
BCQ-Lag	Reward ↑	11.27±0.24	11.18±0.28	11.03±0.11	12.63±0.56	12.41±0.26	12.18±0.38	11.78
	Cost ↓	6.05±0.64	6.38±0.47	6.80±0.59	4.84±0.51	5.09±1.11	5.30±0.76	5.74
	Cost_rate(%) ↓	0.20±0.10	0.20±0.12	0.28±0.15	0.15±0.08	0.16±0.08	0.16±0.06	0.19
CCRR	Reward ↑	11.67±0.24	11.44±0.22	11.34±0.17	12.57±0.80	12.30±0.61	12.21±0.05	11.92
	Cost ↓	6.11±2.34	6.45±2.64	6.80±1.48	4.16±0.85	4.89±1.04	5.26±0.79	5.61
	Cost_rate(%) ↓	0.21±0.13	0.27±0.16	0.31±0.17	0.15±0.07	0.15±0.09	0.16±0.12	0.21
CPQ-Lag	Reward ↑	11.18±2.53	10.97±0.26	10.90±0.43	14.66±0.44	14.17±0.53	13.94±0.42	12.64
	Cost ↓	7.94±1.46	6.24±1.38	6.60±2.00	8.59±1.87	7.28±1.10	7.97±0.69	7.44
	Cost_rate ↓	0.32±0.18	0.22±1.41	0.26±0.18	0.38±0.13	0.24±0.15	0.26±0.12	0.28
POCE	Reward ↑	**12.74±0.26**	**12.62±0.13**	**12.54±0.19**	**13.58±0.78**	**13.45±0.51**	13.41±0.64	**13.06**
	Cost ↓	4.37±0.78	4.76±0.89	4.93±0.83	3.43±0.65	4.56±0.62	5.00±0.43	4.51
	Cost_rate(%) ↓	0.09±0.04	0.12±0.05	0.14±0.06	0.04±0.01	0.06±0.05	0.11±0.04	0.09

spectively, enabling the optimization of multi-constrained policies in an offline setting. Similarly, we introduce two Lagrange multipliers based on CRR [43] to obtain an improved multi-constrained offline RL algorithm CCRR. CPQ-Lag is an extension of the constrained offline RL algorithm CPQ [44] that introduces additional Lagrange multipliers to handle the state-wise cost Q-values, thus enabling the implementation of a multi-constraint offline RL method.

5.2. Performance Comparison Experiment

Performance on various tasks and behavioral samples. To evaluate the performance of the POCE algorithm across various tasks and behavioral samples, we conducted a comprehensive evaluation. This evaluation involved comparing the performance of the POCE with the baseline algorithms in three tasks: *PointGoal*, *CarGoal*, and *AntVelocity*. Furthermore, we also examined the algorithm's performance across samples with different safety factors φ. Fig. 2 illustrates the reward, cumulative cost, and state cost violation rate curves for the baseline algorithm and the POCE algorithm in the *PointGoal* task. From the illustrated results, it can be observed that compared to the baseline algorithm, POCE not only constrains the cumulative cost within the cost threshold range but also achieves a lower state-wise cost violation rate. Furthermore, it provides competitive rewards return. The results presented in Table 1 demonstrate that the POCE algorithm consistently ensures that the policy satisfies cumulative cost constraints across samples with different safety factors while also delivering competitive reward returns. Notably, in the *PointGoal* task, the POCE outperforms the baseline algorithms in terms of both reward and cost-effectiveness. The above results indicate that the POCE algorithm can consistently ensure policy compliance with cumulative cost constraints across samples with different behaviors while providing competitive reward returns and state-wise cost management.

Performance on different state-wise cost thresholds. To evaluate the performance of the POCE algorithm under different state-wise cost thresholds, we conducted comparative experiments between the POCE algorithm and the baseline algorithm at multiple state-wise cost thresholds. The results in Table 2 demonstrate that POCE provides competitive reward returns while ensuring the satisfaction of cumulative cost constraints under different state-wise cost thresholds. Moreover, the mean values of reward and cumulative costs for both tasks are superior to those of the baseline algorithm. Additionally, from the results in the table, it is evident that lowering the state-wise cost threshold leads to a decrease in cumulative costs. This observation is consistent with the viewpoint expressed in Remark 4.2. The analysis of the above results reveals that the POCE algorithm ensures compliance with cost constraints for cumulative costs under different state-wise cost thresholds, while also providing competitive rewards and state-wise cost violation rates.

5.3. Ablation Experiment

The parameters safety factors φ and state-wise cost thresholds $\bar{D}_i$. To assess the sensitivity of the POCE algorithm to different safety factors φ and state-wise cost thresholds $\bar{c}_i$, we tested the performance of the POCE algorithm at different state-wise cost thresholds $\bar{c}_i$ using samples with varying safety factors φ in the *PointGoal* task. Fig. 3 displays the mean performance of the POCE algorithm at different state-wise cost thresholds with samples of varying safety factors in the *PointGoal* task. From the results in the figure, it can be observed that higher safety factors in the sample or higher state-wise cost thresholds result in larger reward returns. Conversely, lower safety factors in the sample or lower state-wise cost thresholds lead a higher state-wise cost violation rates. Additionally, higher safety factors φ in the sample data or lower state-wise cost thresholds $\bar{c}_i$ result in lower cumulative costs. From the above analysis, it is evident that higher safety factors in the sample data lead to higher rewards, along with lower cumulative costs and

Table 2. The performance of the baseline and POCE algorithms is assessed under various state-wise cost thresholds. The results from experiments involving 20 episodes are conducted with three random seeds. The data with a shaded background indicated the satisfaction of cumulative cost constraints, while the data highlighted in bold represents the highest reward return obtained under the satisfaction of cumulative cost constraints. The safety factor for the experiment is $\varphi = 0.8$, and the cumulative cost threshold is consistent at $\bar{c}_i = 5$.

Method	Metrics	PointGoal			CarGoal			Mean
		$\bar{D}_i$=0.8	$\bar{D}_i$=0.6	$\bar{D}_i$=0.4	$\bar{D}_i$=0.8	$\bar{D}_i$=0.6	$\bar{D}_i$=0.4	
BCQ-Lag	Reward ↑	11.27±0.24	11.13±0.34	11.06±0.20	12.63±0.56	12.40±0.46	11.28±1.17	11.63
	Cost ↓	6.05±0.64	5.85±1.97	4.65±0.94	4.84±0.51	4.08±0.46	3.95±0.56	4.90
	Cost_rate(%) ↓	0.20±0.10	1.01±0.49	1.48±0.65	0.15±0.08	0.38±0.18	1.03±0.82	0.71
CCRR	Reward ↑	11.67±0.24	11.50±0.20	11.30±0.20	12.57±0.80	12.23±0.73	12.16±0.76	11.91
	Cost ↓	6.11±2.34	5.95±2.37	5.29±2.37	4.16±0.85	4.19±0.85	4.04±0.92	4.96
	Cost_rate ↓	0.21±0.13	1.29±1.08	1.72±1.13	0.15±0.07	0.47±0.30	1.10±0.68	0.82
CPQ-Lag	Reward ↑	11.18±2.53	10.97±1.64	10.59±1.26	14.66±0.44	14.11±0.31	13.73±0.25	12.54
	Cost ↓	7.94±1.46	7.27±2.61	6.65±3.31	8.59±1.87	7.92±1.03	5.99±1.18	7.39
	Cost_rate (%) ↓	0.32±0.18	2.13±1.57	2.98±1.38	0.38±0.13	0.92±0.52	2.03±1.47	1.46
POCE	Reward ↑	**12.74±0.26**	**12.58±0.16**	**12.49±0.24**	**13.58±0.78**	**13.41±0.50**	**13.28±0.51**	**13.01**
	Cost ↓	4.37±0.78	4.15±0.37	4.07±0.67	3.43±0.65	3.33±0.56	3.24±0.75	3.77
	Cost_rate (%) ↓	0.09±0.04	0.24±0.16	0.98±0.75	0.04±0.01	0.20±0.09	0.80±0.35	0.39

cost violation rates. On the other hand, lower state-wise cost thresholds result in higher cost violation rates, while leading to lower rewards and cumulative costs.

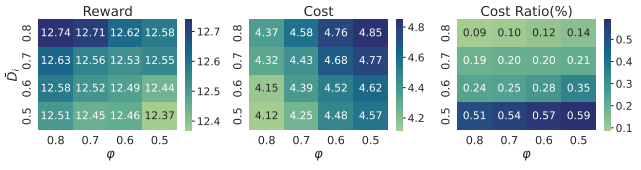

Figure 3. The mean values of rewards, cumulative costs, and stat-wise cost violation rates for the POCE algorithm under varying safety factors and state-wise cost thresholds in the *PointGoal* task. The values shown in the figure represent the means of 20 episodes tested independently with three different random seeds.

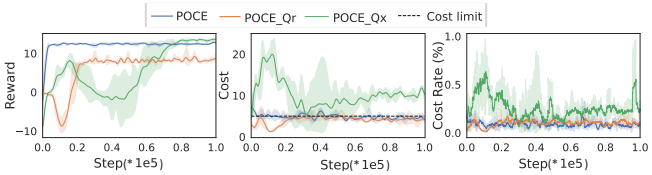

Figure 4. The reward, cumulative cost, and state-wise cost violation rate curves of the POCE algorithm and the POCE algorithm with conservative Q-value estimation removed in the *PointGoal* task. The shadowed curves represent the mean and variance of the test results for three different random seeds.

Conservative estimation of rewards and costs. We evaluate the impact of conservative estimation of POCE algorithm by removing the conservative estimation methods for reward and cost Q-values. POCE-Qr represents the POCE algorithm with the removal of conservative reward Q-values, while POCE-Qx represents the POCE algorithm with the removal of conservative cumulative cost Q-values and state cost Q-values based on conditional Bellman estimation. Fig. 4 displays the reward, cumulative cost, and state-wise cost violation rate curves for the POCE algorithm and the POCE algorithm with conservative Q-value estimation removed in the *PointGoal* task. From the figure, it can be observed that the reward of POCE-Qr is significantly lower compared to POCE, while the reward of POCE-Qx is slightly higher than POCE. However, POCE-Qx fails to satisfy the cumulative cost constraint, and the state-wise cost violation rate also noticeably increases. This indicates that conservative estimation of reward Q-values can improve the reward performance of the POCE algorithm, while conservative estimation of cost Q-values based on conditional Bellman implementation can prevent underestimation of Q-values for action-state pairs, thereby ensuring safety.

6. Conclusion

In this work, we propose a novel policy optimization algorithm with conservative estimation for multi-constrained offline reinforcement learning. Concretely, we first rethink the requirements of constraint RL in real-world applications and redefine the objectives of multi-constrained offline RL tasks by introducing the MMDP. Then, we present a primal policy optimization method to address the multi-constrained optimization problem. Additionally, we propose the conditional Bellman operator to achieve a conservative estimation of cumulative cost and state-wise cost. Finally, extensive experiments demonstrate that the POCE algorithm provides competitive performance, particularly in terms of safety.

Acknowledgment: This work is supported by STI 2030-Major Projects (No.2021ZD0201405), in part by the National Natural Science Foundation of China (No. 62372329), in part by the National Key Research and Development Program of China (No.2021YFB2501104), in part by Shanghai Rising Star Program (No.21QC1400900), in part by Tongji-Qomolo Autonomous Driving Commercial Vehicle Joint Lab Project, and in part by Xiaomi Young Talents Program. We thank Long Yang and Yiqin Yang for the insightful discussion.

References

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017. 2

[2] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114. PMLR, 2020. 1

[3] Eitan Altman. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999. 2

[4] Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[5] Xueying Bai, Jian Guan, and Hongning Wang. A model-based reinforcement learning with adversarial training for online recommendation. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 1

[6] Steven Bohez, Abbas Abdolmaleki, Michael Neunert, Jonas Buchli, Nicolas Heess, and Raia Hadsell. Value constrained model-free continuous control. *arXiv preprint arXiv:1902.04623*, 2019. 1

[7] Xu Chen, Yali Du, Long Xia, and Jun Wang. Reinforcement recommendation with user multi-aspect preference. In *Proceedings of the Web Conference 2021*, pages 425–435, 2021. 1

[8] 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. 2

[9] 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. 2

[10] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019. 6

[11] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015. 2

[12] 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. 1

[13] Jiayi Guan, Guang Chen, Jiaming Ji, Long Yang, Ao Zhou, Zhijun Li, and changjun jiang. VOCE: Variational optimization with conservative estimation for offline safe reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 6

[14] Jiayi Guan, Shangding Gu, Zhijun Li, Jing Hou, Yiqin Yang, Guang Chen, and Changjun Jiang. Uac: Offline reinforcement learning with uncertain action constraint. *IEEE Transactions on Cognitive and Developmental Systems*, 2023. 1

[15] Jing Hou, Guang Chen, Zhijun Li, Wei He, Shangding Gu, Alois Knoll, and Changjun Jiang. Hybrid residual multiexpert reinforcement learning for spatial scheduling of high-density parking lots. *IEEE Transactions on Cybernetics*, pages 1–13, 2023. 1

[16] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1

[17] Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13739–13748, 2022. 1

[18] Jiaming Ji, Jiayi Zhou, Borong Zhang, Juntao Dai, Xuehai Pan, Ruiyang Sun, Weidong Huang, Yiran Geng, Mickel Liu, and Yaodong Yang. Omnisafe: An infrastructure for accelerating safe reinforcement learning research. *arXiv preprint arXiv:2305.09304*, 2023. 6

[19] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021. 1

[20] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020. 5

[21] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019. 1

[22] Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pages 6120–6130. PMLR, 2021. 1, 2, 3

[23] Qingkai Liang, Fanyu Que, and Eytan Modiano. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480*, 2018. 4

[24] Yongshuai Liu, Jiaxin Ding, and Xin Liu. Ipo: Interior-point policy optimization under constraints. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4940–4947, 2020. 1, 2

[25] Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, and Ding Zhao. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*, pages 13644–13668. PMLR, 2022. 2

[26] 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. 6

[27] Zuxin Liu, Zijian Guo, Yihang Yao, Zhepeng Cen, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constrained decision transformer for offline safe reinforcement learning. *arXiv preprint arXiv:2302.07351*, 2023. 2

[28] Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. Mildly conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 1711–1724, 2022. 5

[29] Tatsuya Matsushima, Hiroki Furuta, Yutaka Matsuo, Ofir Nachum, and Shixiang Gu. Deployment-efficient reinforcement learning via model-based offline optimization. In *International Conference on Learning Representations*, 2021. 1

[30] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. *Advances in neural information processing systems*, 29, 2016. 1

[31] Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 68(3):1321–1336, 2022. 1

[32] Nicholas Polosky, Bruno C Da Silva, Madalina Fiterau, and Jithin Jagannath. Constrained offline policy optimization. In *International Conference on Machine Learning*, pages 17801–17810. PMLR, 2022. 1, 2, 3

[33] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7077–7087, 2021. 1

[34] LA Prashanth. Policy gradients for cvar-constrained mdps. In *International Conference on Algorithmic Learning Theory*, pages 155–169. Springer, 2014. 2

[35] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pages 387–395. Pmlr, 2014. 4

[36] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143. PMLR, 2020. 1

[37] Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019. 2, 4

[38] Garrett Thomas, Yuping Luo, and Tengyu Ma. Safe reinforcement learning by imagining the near future. *Advances in Neural Information Processing Systems*, 34:13859–13869, 2021. 1, 2

[39] Marin Toromanoff, Emilie Wirbel, and Fabien Moutarde. End-to-end model-free reinforcement learning for urban driving using implicit affordances. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7153–7162, 2020. 1

[40] Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806. PMLR, 2020. 2

[41] 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*, 2018. 2

[42] Xudong Wang, Long Lian, and Stella X Yu. Unsupervised visual attention and invariance for reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6677–6687, 2021. 1

[43] Ziyu Wang, Alexander Novikov, Konrad Zolna, Josh S Merel, Jost Tobias Springenberg, Scott E Reed, Bobak Shahriari, Noah Siegel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. *Advances in Neural Information Processing Systems*, 33:7768–7778, 2020. 7

[44] Haoran Xu, Xianyuan Zhan, and Xiangyu Zhu. Constraints penalized q-learning for safe offline reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8753–8760, 2022. 1, 2, 7

[45] 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. 4

[46] Xintao Yan, Zhengxia Zou, Shuo Feng, Haojie Zhu, Haowei Sun, and Henry X Liu. Learning naturalistic driving environment with statistical realism. *Nature Communications*, 14(1):2037, 2023. 1

[47] Long Yang, Jiaming Ji, Juntao Dai, Linrui Zhang, Binbin Zhou, Pengfei Li, Yaodong Yang, and Gang Pan. Constrained update projection approach to safe policy optimization. *arXiv preprint arXiv:2209.07089*, 2022. 2

[48] Qisong Yang, Thiago D Simão, Simon H Tindemans, and Matthijs TJ Spaan. Wcsac: Worst-case soft actor critic for safety-constrained reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10639–10646, 2021. 2

[49] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Projection-based constrained policy optimization. *arXiv preprint arXiv:2010.03152*, 2020. 2

[50] Yijun Yang, Jing Jiang, Tianyi Zhou, Jie Ma, and Yuhui Shi. Pareto policy pool for model-based offline reinforcement learning. In *International Conference on Learning Representations*, 2021. 1

[51] Yijun Yang, Tianyi Zhou, Jing Jiang, Guodong Long, and Yuhui Shi. Continual task allocation in meta-policy network via sparse prompting. In *International Conference on Machine Learning*, pages 39623–39638. PMLR, 2023. 2

[52] Chengyang Ying, Xinning Zhou, Hang Su, Dong Yan, Ning Chen, and Jun Zhu. Towards safe reinforcement learning via constraining conditional value-at-risk. *arXiv preprint arXiv:2206.04436*, 2022. 2

[53] Ming Yu, Zhuoran Yang, Mladen Kolar, and Zhaoran Wang. Convergent policy optimization for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[54] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020. 1

[55] Linrui Zhang, Li Shen, Long Yang, Shixiang Chen, Bo Yuan, Xueqian Wang, and Dacheng Tao. Penalized proximal policy optimization for safe reinforcement learning. *arXiv preprint arXiv:2205.11814*, 2022. 2

[56] Linrui Zhang, Qin Zhang, Li Shen, Bo Yuan, Xueqian Wang, and Dacheng Tao. Evaluating model-free reinforcement learning toward safety-critical tasks. *Proceedings of the*

AAAI Conference on Artificial Intelligence, 37(12):15313–15321, 2023. 2

[57] Qin Zhang, Linrui Zhang, Haoran Xu, Li Shen, Bowen Wang, Yongzhe Chang, Xueqian Wang, Bo Yuan, and Dacheng Tao. Saformer: A conditional sequence modeling approach to offline safe reinforcement learning. *arXiv preprint arXiv:2301.12203*, 2023. 2

[58] Ruiqi Zhang, Jing Hou, Guang Chen, Zhijun Li, Jianxiao Chen, and Alois Knoll. Residual policy learning facilitates efficient model-free autonomous racing. *IEEE Robotics and Automation Letters*, 7(4):11625–11632, 2022. 1

[59] Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. *Advances in Neural Information Processing Systems*, 33:15338–15349, 2020. 2

[60] Weiye Zhao, Rui Chen, Yifan Sun, Tianhao Wei, and Changliu Liu. State-wise constrained policy optimization. *arXiv preprint arXiv:2306.12594*, 2023. 2

[61] Weiye Zhao, Tairan He, Rui Chen, Tianhao Wei, and Changliu Liu. State-wise safe reinforcement learning: A survey. *arXiv preprint arXiv:2302.03122*, 2023. 2, 3, 4