Constraints Penalized Q-learning for Safe Offline Reinforcement Learning

Haoran Xu1,2,3, Xianyuan Zhan4*, Xiangyu Zhu2,3

1 School of Computer Science and Technology, Xidian University, Xi’an, China
2 JD iCity, JD Technology, Beijing, China
3 JD Intelligent Cities Research
4 Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
{ryanxhr, zhanxianyuan, zackxiangyu}@gmail.com

Abstract

We study the problem of safe offline reinforcement learning (RL), the goal is to learn a policy that maximizes long-term reward while satisfying safety constraints given only offline data, without further interaction with the environment. This problem is more appealing for real world RL applications, in which data collection is costly or dangerous. Enforcing constraint satisfaction is non-trivial, especially in offline settings, as there is a potential large discrepancy between the policy distribution and the data distribution, causing errors in estimating the value of safety constraints. We show that naive approaches that combine techniques from safe RL and offline RL can only learn sub-optimal solutions. We thus develop a simple yet effective algorithm, Constraints Penalized Q-Learning (CPQ), to solve the problem. Our method admits the use of data generated by mixed behavior policies. We present a theoretical analysis and demonstrate empirically that our approach can learn robustly across a variety of benchmark control tasks, outperforming several baselines.

Introduction

Reinforcement Learning (RL) has achieved great success in solving complex tasks, including games (Mnih et al. 2013; Silver et al. 2017), and robotics (Levine et al. 2016). However, most RL algorithms learn good policies only after millions of trials and errors in simulation environments. Consider real-world scenarios (e.g. self-driving cars, industrial control systems), where we only have a batch of pre-collected data (non-optimal), including some unsafe attempts (e.g. high-speed collisions in self-driving cars), no further active online data collection is allowed. The question then arises: how can we derive an effective policy from these offline data while satisfying safety constraints?

Safe RL is usually modeled as a Constrained Markov Decision Process (CMDP) (Altman 1999). There are typically two kinds of constraints: hard constraints and soft constraints. Hard constraints require no constraints violation at each time step of the trajectory, while soft constraints require the policy to satisfy constraints in expectation throughout the whole trajectory. In this work, we focus on the soft constraints. There is a branch of related work for safe RL, with the main focus on safe exploration (Chow et al. 2017; Achiam et al. 2017; Tessler, Mankowitz, and Mannor 2018). However, none of these algorithms are off-policy and cannot be used in offline settings. Although one recent study aims towards batch policy learning under constraints (Le, Voloshin, and Yue 2019), this method assumes sufficient exploration of the data collection policy. This requirement usually does not hold in real-world scenarios, especially in high-dimensional continuous control tasks.

The key challenge is how to evaluate constraint violations accurately while maximizing reward effectively. Typically, this needs to roll out the policy in the environment and evaluate constraint values by on-policy samples (Tessler, Mankowitz, and Mannor 2018). However, it is impossible in the offline setting, as we only have access to samples in the offline dataset. Evaluating constraint values from offline data is non-trivial, it will encounter serious issues when the evaluated policy lies outside of the dataset distribution. In value-based RL approaches, this may introduce errors in the Q-function backup and it is impossible to collect online data to correct such errors. The problem will be further exacerbated when the offline dataset is generated by multiple conflicting behavior policies, as the policy may be biased to be unsafe or sub-optimal.

One can use an extra cost critic (like the reward critic) to learn the constraint values, and use a divergence penalty to control the deviation between the learned policy and the dataset distribution. However, we show that this naive method is too conservative and will lead to a sub-optimal solution. Our primary contribution is presenting a new algorithm, Constraints Penalized Q-Learning (CPQ), to address the above challenges. The intuition is that besides those original unsafe actions, we additionally make those actions that are out of the data distribution unsafe. To accomplish this, we modify the Bellman update of reward critic to penalize those state action pairs that are unsafe. CPQ does not use an explicit policy constraint and will not be restricted by the density of the dataset distribution, it admits the use of datasets generated by mixed behavior policies. We also provide a theoretical error bound analysis of CPQ under mild assumptions. Through systematic experiments, we show that our algorithm can learn robustly to maximize rewards while successfully satisfying safety constraints, outperform all baselines in benchmark continuous control tasks.

*Corresponding author
Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Safe RL can be defined as the process of learning policies that maximizes long-term rewards while ensuring safety constraints. When Markov transition probability is known, a straightforward approach is based on linear programming (Altman 1999). In model-free settings, Lagrangian-based methods (Chow et al. 2017; Tessler, Mankowitz, and Mannor 2018) augment the standard expected reward objective with a penalty of constraint violation and solve the resulting problem with a learnable Lagrangian multiplier. However, Lagrangian-based policy can only asymptotically satisfy the constraint and makes no safety guarantee during the training process when interaction with the real-world environment is required. Constrained policy optimization (CPO) (Achiam et al. 2017) extends trust-region optimization (Schulman et al. 2015), which can satisfy constraints during training, but the computational expense dramatically increases with multiple constraints. There are some other approaches designed for convex constraints (Miryoosefi et al. 2019) or hard constraints (Dalal et al. 2018; Satija, Amortila, and Pineau 2020). However, all of these algorithms are on-policy, thus cannot be applied to the offline setting. Constrained Batch Policy Learning (CBPL) (Le, Voloshin, and Yue 2019) considers safe policy learning offline, it uses Fitted Q Evaluation (FQE) to evaluate the safe constraints and learn the policy by Fitted Q Iteration (FQI), through a game-theoretic framework.

Off-Policy RL algorithms based on dynamic programming maintain a parametric Q-function \( Q_\phi(s, a) \). Q-learning methods (Watkins and Dayan 1992) train the Q-function by iteratively applying the Bellman optimality operator \( \mathcal{T}^\pi Q(s, a) := r(s, a) + \gamma \max_{a' \in A} Q(s', a') \). In an actor-critic algorithm, the Q-function (critic) is trained by iterating the Bellman evaluation operator \( \mathcal{T}^\pi Q = r + \gamma P^\pi Q \), where \( P^\pi \) is the transition matrix coupled with the policy: \( P^\pi Q(s, a) = \mathbb{E}_{s' \sim T(s, a) \sim \pi} Q(s', a') \), and a separate policy is trained to maximize the expected Q-value. Since the replay buffer typically does not contain all possible transitions \((s, a, s')\), the policy evaluation step actually uses an empirical Bellman operator that only back up a single sample \( s' \). Notice that \( \pi \) is trained to maximize Q-values, it may be biased towards out-of-distribution (OOD) actions with erroneously high Q-values. In standard (online) RL, such errors can be corrected by interacting with the environment and observing its actual value.

In our problem, we assume no interaction with the environment and only have a batch, offline dataset \( B = (s, a, r, s', c, (s, a), (c, s, a)) \), generated by following unknown arbitrary behavior policies. Note that these behavior policies may generate trajectories that violate safety constraints. We use \( \pi_\beta \) to represent the empirical behavior policy of the dataset, formally, \( \pi_\beta(a_0 | s_0) := \frac{\sum_{s, a, r, s'} I[s = s_0, a = a_0]}{\sum_{s, a, r, s'} I[s = s_0]}, \) for all state \( s_0 \in B \). We use \( \mu^\beta(s) \) to represent the discounted marginal state-distribution of \( \pi_\beta(a | s) \), thus the dataset \( B \) is sampled from \( \mu^\beta(s) \pi_\beta(a | s) \). The goal of safe offline learning is to learn a policy \( \pi \) from \( B \) that maximizes the cumulative reward while satisfying the cumulative cost constraint, denoted as

\[ \max_\pi \quad R(\pi) \]
\[ \text{s.t.} \quad C(\pi) \leq l \]

where \( l \) is the safe constraint limit (a known constant).

| Related Work |
|---------------|
| Safe Reinforcement Learning |
| Offline RL (also known as batch RL (Lange, Gabel, and Riedmiller 2012) or fully off-policy RL) considers the problem of learning policies from offline data without interaction with the environment. One major challenge of offline RL is the distributional shift problem (Levine et al. 2020), which incurs when the policy distribution deviates largely from the data distribution. Although off-policy RL methods (Mnih et al. 2013; Lillicrap et al. 2016) are naturally designed for tackling this problem, they typically fail to learn solely from fixed offline data and often require a growing batch of online samples for good performance. Most recent methods attempted to solve this problem by constraining the learned policy to be “close” to the behavior policy. BCQ (Fujimoto, Meger, and Precup 2019) learns a generative model for the behavior policy and adds small perturbations to it to stay close to the data distribution while maximizing the reward. Some other works use divergence penalties (such as KL divergence in BRAC (Wu, Tucker, and Nachum 2019) or maximum mean discrepancy (MMD) in BEAR (Kumar et al. 2019)) instead of perturbing actions. CQL (Kumar et al. 2020) uses an implicit Q-value constraint between the learned policy and dataset samples, which avoids estimating the behavior policy. The distributional shift problem can also be solved by model-based RL through a pessimistic MDP framework (Yu et al. 2020; Kidambi et al. 2020; Zhan et al. 2021) or by constrained offline model-based control (Argenson and Dulac-Arnold 2021; Zhan, Zhu, and Xu 2021). |

| Preliminary |
| A Constrained Markov Decision Process (CMDP) is represented by a tuple \((\mathcal{S}, A, r, c, P, \gamma, \eta)\), where \( \mathcal{S} \subset \mathbb{R}^n \) is the closed and bounded state space and \( A \subset \mathbb{R}^m \) is the action space. Let \( r : \mathcal{S} \times \mathcal{A} \rightarrow [0, R] \) and \( c : \mathcal{S} \times \mathcal{A} \rightarrow [0, C] \) denote the reward and cost function, bounded by \( R \) and \( C \). Let \( P : \mathcal{S} \times A \times \mathcal{S} \rightarrow [0, 1] \) denote the (unknown) transition probability function that maps state-action pairs to a distribution over the next state. Let \( \eta \) denote the initial state distribution. And finally, let \( \gamma \in [0, 1) \) denote the discount factor for future reward and cost. A policy \( \pi : \mathcal{S} \rightarrow \mathcal{P}(A) \) corresponds to a map from states to a probability distribution over actions. Specifically, \( \pi(a | s) \) denotes the probability of taking action \( a \) in state \( s \). In this work, we consider parametrized policies (e.g., neural networks), we may use \( \pi_\theta \) to emphasize its dependence on parameter \( \theta \). The cumulative reward under policy \( \pi \) is denoted as \( R(\pi) = \mathbb{E}_{t \sim \pi} \sum_{i=0}^\infty \gamma^i r(s_t, a_t) \), where \( \pi = (a_0, a_0, s_1, a_1, \ldots) \) is a trajectory and \( t \sim \tau \) means the distribution over trajectories is induced by policy \( \tau \). Similarly, the cumulative cost takes the form as \( C(\pi) = \mathbb{E}_{t \sim \pi} \sum_{i=0}^\infty \gamma^i c(s_t, a_t) \).

Off-policy RL algorithms based on dynamic programming maintain a parametric Q-function \( Q_\phi(s, a) \). Q-learning methods (Watkins and Dayan 1992) train the Q-function by iteratively applying the Bellman optimality operator \( \mathcal{T}^\pi Q(s, a) := r(s, a) + \gamma \max_{a' \in A} Q(s', a') \). In an actor-critic algorithm, the Q-function (critic) is trained by iterating the Bellman evaluation operator \( \mathcal{T}^\pi Q = r + \gamma P^\pi Q \), where \( P^\pi \) is the transition matrix coupled with the policy: \( P^\pi Q(s, a) = \mathbb{E}_{s' \sim T(s, a) \sim \pi} Q(s', a') \), and a separate policy is trained to maximize the expected Q-value. Since the replay buffer typically does not contain all possible transitions \((s, a, s')\), the policy evaluation step actually uses an empirical Bellman operator that only back up a single sample \( s' \). Notice that \( \pi \) is trained to maximize Q-values, it may be biased towards out-of-distribution (OOD) actions with erroneously high Q-values. In standard (online) RL, such errors can be corrected by interacting with the environment and observing its actual value. |
An Naïve Approach

An naïve approach to solve safe offline RL is combining techniques from safe RL and offline RL. For example, we could train an additional cost critic Q-network (to get the value of cumulative cost like in (Liang, Que, and Modiano 2018; Ha et al. 2020)), along with a divergence constraint to prevent large distributional shift from $\beta$. Formally, we update both the reward and cost critic Q-network by the empirical Bellman evaluation operator $T^\pi$ for $(s, a, s', r, c) \sim B$:

$$Q_r(s, a) = r + \gamma E_{a' \sim \pi(s')} [Q_r(s', a')]$$

$$Q_c(s, a) = c + \gamma E_{a' \sim \pi(s')} [Q_c(s', a')]$$

and then the policy can be derived by solving the following optimization problem:

$$\pi_\theta := \max_{\pi \in \Delta_{|s|}} E_{s \sim B, a \sim \pi(s)} [Q_r(s, a)]$$

s.t. $E_{s \sim B, a \sim \pi(s)} [Q_c(s, a)] \leq \ell$ \hspace{1cm} (Constraint 1)

$D(\pi, \pi_\theta) \leq \xi$ \hspace{1cm} (Constraint 2)

$D$ can be any off-the-shelf divergence metrics (e.g., KL divergence or MMD distance) and $\xi$ is an approximately chosen small value. We can convert the constrained optimization problem to an unconstrained form by using the Lagrangian relaxation procedure, and solve it by dual gradient descent.

However, we argue that in this approach, Constraint 1 and Constraint 2 may not be satisfied simultaneously. Suppose $B$ consists of transitions from both safe and unsafe policies. When the policy $\pi$ satisfies Constraint 2, it will match the density of the behavior policy distribution, when the behavior policy distribution contains some unsafe actions, the resulting policy may violate Constraint 1. One may consider to subtract transitions of the safe policy from $B$ to construct a new "safe dataset", and only use it for training. Although in this case, Constraint 1 and Constraint 2 can be both satisfied, however, the missing of high-reward transitions will make the resulting policy sub-optimal. In principle, by carefully "stitching" together transitions from both safe and unsafe policies, the policies ought to produce trajectories with maximized cumulative reward while still satisfying safety constraints.

Constraints Penalized Q-Learning

In this section, we introduce our method, Constraints Penalized Q-Learning (CPQ), a simple yet effective algorithm for safe offline RL. The key idea is to make OOD actions "unsafe" and update the reward critic using only state-action pairs that are "safe". CPQ avoids explicit policy constraints, it involves the following three steps:

**Step 1:** We first make Qc-values of OOD actions larger than the safe constraint limit, we accomplish this by adding an additional term to the original objective of Bellman evaluation error, yielding a new objective:

$$\min_{Q_c} E_{s,a,s' \sim B} [(Q_c - T^\pi Q_c)^2] - \alpha E_{s \sim B, a \sim \nu} [Q_c(s, a)]$$

Aside from the standard Bellman evaluation error term, Equation (1) also maximizes the Qc-values at all states in the dataset $B$, for those actions induced from the distribution $\nu$.

Intuitively, if we choose $\nu$ to be a distribution that generates OOD actions, those OOD actions’ Qc-values will be pushed up. Note that the Qc-values of in-distribution actions would be pushed down to obey the Bellman backup by the standard Bellman error term. Therefore, with an appropriate weight $\alpha$, we will only overestimate Qc-values of OOD actions, while keep them unchanged for in-distribution actions.

The remaining question is how to get the distribution $\nu$ that generates OOD actions. We avoid this hard problem by performing the OOD detection (Ren et al. 2019; Liu et al. 2020). As the policy $\pi$ is trained to maximize the reward critic, we only need to ensure that actions sampled by $\pi$ are not OOD. To do so, we pretrain the Conditional Variational Autoencoder (CVAE) to model the behavior policy of the dataset and utilize the latent space to do OOD detection. More specifically, we train the state-conditional VAE based on the following evidence lower bound (ELBO) objective on the log-likelihood of the dataset:

$$\max_{\omega_1, \omega_2} \mathbb{E}_{z \sim q_{\omega_2}} [\log p_{\omega_1}(s, z)] - \beta D_{KL} [q_{\omega_2}(z|s, a) \| p_{\omega_1}(z)]$$ \hspace{1cm} (2)

The first term represents the reconstruction loss and the second term is the KL-divergence between the encoder output and the prior of $z$. Note that action $a$ at state $s$ will have a high probability under the behavior data distribution if the value of $z \sim q_{\omega_1}(s, a)$ has a high probability under the prior $p(z)$. Since $p(z)$ is set to be $\mathcal{N}(0, 1)$, we can let $\nu(s) = a$ if $D_{KL}[q_{\omega_2}(z|s, a) \| \mathcal{N}(0, 1)] \geq d$, by introducing a hyperparameter $d$ to control the threshold. Previous works (Fujimoto, Meger, and Precup 2019; Kumar et al. 2019) also use CVAE, but they use it to sample actions and compute the value of divergence metrics, which is different from our usage.

**Step 2:** In Step 1, the cost critic learned by CPQ is somewhat "distorted", i.e., Qc-values of OOD actions are likely to be larger than their true values and extrapolate to actions near the boundary of in-distribution actions. In preliminary experiments, we found it did not work well when using the distorted cost critic to update the policy by dual gradient descent (i.e., $\max_c Q_r - Q_c^\pi$). Fortunately, Qr-values in Step 1 remain unchanged, so we can update the policy by only maximizing Qr-values. We modify the reward critic’s Bellman update to only backup from state action pairs that are both constraint safe and in-distribution safe, this is accomplished by multiplying $Q_r(s', a')$ by an indicator. We define the empirical Constraints Penalized Bellman operator $T^\pi_r$ for $(s, a, s', r, c) \sim B$, as

$$T^\pi_r Q_r(s, a) = r + \gamma E_{a' \sim \pi} [\mathbb{I} (Q_r(s', a') \leq \ell) Q_r(s', a')]$$

where $\mathbb{I}$ is the indicator function. It can be shown that $T^\pi_r$ reduces the update from those unsafe state action pairs by using a pessimistic estimate of 0 to those pairs. Given the offline dataset $B$, we update the reward critic by minimizing the mean-square error (MSE) as

$$\min_{Q_r} E_{s,a,s' \sim B} [(Q_r(s, a) - T^\pi_r Q_r(s, a))^2]$$ \hspace{1cm} (3)

**Step 3:** Finally, in the policy improvement step, to ensure the final policy is safe, CPQ applies the indicator to the computed
state-action values before performing maximization:

$$\pi_\theta := \max_{\pi \in \mathcal{A}_{[2]}} \mathbb{E}_{s \sim B} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[ \mathbf{1} \left( Q_c(s, a) \leq l \right) Q_r(s, a) \right]$$  \hspace{1cm} (4)

**Connections to CQL**  CQL adds two penalty terms to the standard Bellman error term, the first term is minimizing Qr-values of actions from the learned policy, the second term is maximizing Qr-values of actions from the dataset. CQL makes the Qr-value landscape to have higher values in the area of data distribution than in the area of the policy distribution, since CQL only considers reward maximization, it gets rid of the bad impact of OOD actions. In our problem, the policy is trained to both maximize the reward and satisfy safety constraints. We cannot simply follow CQL by maximizing Qc-values of actions from the policy and minimizing Qc-values of actions from the dataset. Maximizing Qc-values of actions from the policy will deteriorate the performance when the policy outputs in-distribution actions. Hence, we detect the actions output by the policy and only make Qc-values of those OOD actions large.

**Practical Considerations** To reduce the number of hyper-parameters, we can automatically tune \( \alpha \) as shown below:

$$\min_{Q_c} \max_{\alpha \geq 0} \mathbb{E}_{s,a,a' \sim B} \left[ \left( Q_c - \mathbb{E}_{s,a',s'} \left[ Q_c(s, a') \right] \right)^2 \right]$$

$$- \alpha \left( \mathbb{E}_{s \sim B, a \sim \pi} \left[ Q_c(s, a) \right] - l_c \right)$$ \hspace{1cm} (5)

Equation (5) implies that \( \alpha \) will stop increasing if Qc-values of OOD actions are larger than \( l_c \). The parameter \( l_c \) should be chosen to be larger than the constraint threshold \( l \), in practice, we use \( l_c = 1.5 \times l \) across all tasks. We use \( \beta \)-VAE (Higgins et al. 2017) to learn better disentangled latent space representations compared to the original VAE framework (can be seen as a special case of \( \beta \)-VAE with \( \beta = 1 \)). We sample \( n \) actions from the policy and choose \( \nu \) to be all the actions that violate the latent space threshold \( \delta \). If none of the \( n \) actions violates, Equation (1) is reduced to the original Bellman evaluation error objective. We also adopt the double-Q technique (Fujimoto, Hoof, and Meger 2018) to penalize the uncertainty in value estimations, we select the minimal value of two reward critics when computing the target Qr-values. This trick is not applied to the cost critic, as it will tend to underestimate the Qc-values. Implementation details and hyperparameter choices can be found in Appendix B. The pseudo-code of CPQ is presented in Algorithm 1.

**Analysis**

In this section, we give a theoretical analysis of CPQ, specifically, we prove that we can learn a safe and high-reward policy given only the offline dataset. We first give the notation used for the proof and define what is Out-of-distribution Action Set, then we prove that CPQ can make Qc-values of out-of-distribution actions greater than \( l \) with specific \( \alpha \). Finally we give the error bound of the difference between the Qr-value obtained by iterating Constraints Penalized Bellman operator and the Qr-value of the optimal safe policy \( \pi^* \) that can be learned from the offline dataset.

**Notation** Let \( Q^k \) denotes the true tabular Q-function at iteration \( k \) in the MDP, without any correction. In an iteration, the current tabular Q-function, \( Q^{k+1} \) is related to the previous tabular Q-function iterate \( Q^k \) as: \( Q^{k+1} = \mathbb{E}_{s,a} \left[ Q^k(s, a) \right] \). Let \( Q^k \) denote the \( \nu \)-th Q-function iterate obtained from CPQ. Let \( V^k \) denote the value function as \( V^k := \mathbb{E}_{s,a} \left[ Q^k(s, a) \right] \).

We begin with the definition of Out-of-distribution Action Set:

**Definition 1** (Out-of-distribution Action Set). Given a dataset \( B \), its empirical behavior policy \( \pi_\beta \) and \( \epsilon \in (0, 1) \), we call a set of actions \( A_\epsilon \) (generated by the policy \( \nu \)) as the out-of-distribution action set, if \( \forall s \in B, \forall a \in A_\epsilon, \pi_\beta(a|s) < \epsilon \).

Intuitively, for those out-of-distribution actions (i.e., un-likely to be in the data distribution), \( \nu(a|s) \) will be large while \( \pi_\beta(a|s) \) will be small. In contrast to out-of-distribution actions, in-distribution actions refer to those actions \( a \sim \pi(a|s) \), i.e., have good support in the data distribution. Notice that here we do not care about out-of-distribution states as states used for training are sampled from \( B \). After introducing the out-of-distribution action set, we now show that we can make Qc-values of \( A_\epsilon \) greater than \( l \) with appropriate \( \alpha \) when updating the cost critic by Equation (1).

**Theorem 1.** For any \( \nu(a|s) \) with \( \text{supp} \nu \subset \text{supp} \pi_\beta, \forall s \in B, a \in A_\epsilon, Q^\pi_\beta (the Q-function obtained by iterating Equation (1)) satisfies:

$$Q^\pi_\beta(s, a) = Q^\pi_c(s, a) + \alpha \cdot \frac{1}{2} \left[ (1 - \gamma P^\pi)^{-1} \frac{\nu(s, a)}{\pi_\beta(s, a)} \right]$$

and we can get \( \hat{Q^\pi_\beta}(s, a) \geq l, \forall s \in B, a \in A_\epsilon \), if we choose \( \alpha \geq \max \left[ 2 \epsilon \max_{s,a} (l - Q^\pi_c(s, a)) / (1 - \gamma P^\pi(s, a), 0) \right] \).

---

**Algorithm 1: Constraints Penalized Q-Learning (CPQ)**

**Require:** \( B \), constraint limit \( l \), threshold \( d \).
1. Initialize encoder \( E_{w_1} \) and decoder \( D_{w_2} \).
2. **/VAE Training**
   3. for \( t = 0, 1, \ldots, M \) do
      4. Sample mini-batch of state-action pairs \( (s, a) \sim B \)
      5. Update encoder and decoder by Eq.(2)
   end for
3. **/Policy Training**
   6. Initialize reward critics \( \{ Q_{c_i}(s, a|\phi_{c_i}) \}_{i=1}^l \), cost critic \( Q_c(s, a|\phi_c) \), actor \( \pi_\theta \), Lagrange multiplier \( \alpha \), target networks \( \{ Q^{i'}_c \}_{i=1}^l \) and \( Q^\pi_c \), with \( \phi_{c_i}' \leftarrow \phi_c \) and \( \phi_c' \leftarrow \phi_c \. 
   7. for \( t = 0, 1, \ldots, N \) do
      8. Initialize reward critics \( \{ Q_{c_i}(s, a|\phi_{c_i}) \}_{i=1}^l \)
      9. Sample mini-batch of transitions \( (s, a, r, c, s') \sim B \)
   end for
10. Update actor by Eq.(4) using policy gradient.
11. Update target cost critic: \( \phi_{c_i}' \leftarrow \tau \phi_c + (1 - \tau) \phi_{c_i}' \)
12. Update target reward critics: \( \phi_c' \leftarrow \tau \phi_c + (1 - \tau) \phi_c' \)
13. Update cost critic by Eq.(1) and reward critics by Eq.(3).
14. Update cost critic by Eq.(4) using policy gradient.
15. Update target cost critic: \( \phi'_{c_i} \leftarrow \tau \phi_{c_i} + (1 - \tau) \phi'_{c_i} \)
16. Update target reward critics: \( \phi_c' \leftarrow \tau \phi_c + (1 - \tau) \phi_c' \)
17. end for
Proof. By setting the derivative of Equation (1) to 0, we obtain the following expression for $\hat{Q}_{c}^{k+1}$ in terms of $\hat{Q}_{c}^{k}$,
\[ \forall k, \quad \hat{Q}_{c}^{k+1}(s, a) = T^\pi \hat{Q}_{c}^{k}(s, a) + \frac{\alpha}{2} \cdot \frac{\nu(a|s)}{\pi_\beta(a|s)} \] (6)
Since $\nu(a|s) > 0$, $\alpha > 0$, $\pi_\beta(a|s) > 0$, we observe that at each iteration we enlarge the next Qc-value, i.e., $\hat{Q}_{c}^{k+1} \geq T^\pi \hat{Q}_{c}^{k}$.

Now let’s examine the fixed point of Equation (6) as,
\[ \hat{Q}_{c}^*(s, a) = T^\pi \hat{Q}_{c}^*(s, a) + \frac{\alpha}{2} \cdot \frac{\nu(a|s)}{\pi_\beta(a|s)} \]
\[ = c + \gamma P^\pi \hat{Q}_{c}^*(s, a) + \frac{\alpha}{2} \cdot \frac{\nu(a|s)}{\pi_\beta(a|s)} \]
\[ = Q_{c}^*(s, a)(I - \gamma P^\pi) + \gamma P^\pi \hat{Q}_{c}^*(s, a) + \frac{\alpha}{2} \cdot \frac{\nu(a|s)}{\pi_\beta(a|s)} \]
\[ = Q_{c}^*(s, a) + \gamma P^\pi \left[ \hat{Q}_{c}^*(s, a) - Q_{c}^*(s, a) \right] + \frac{\alpha}{2} \cdot \frac{\nu(a|s)}{\pi_\beta(a|s)} \]

So we can get the relationship between $\hat{Q}_{c}^*$ and the true Qc-value $Q_{c}^*$ as,
\[ \hat{Q}_{c}^*(s, a) = Q_{c}^*(s, a) + \frac{\alpha}{2} \left( I - \gamma P^\pi \right)^{-1} \left[ \frac{\nu(a|s)}{\pi_\beta(a|s)} \right] (s, a) \]

If $Q_{c}^*(s, a) \geq l$, i.e., the true Qc-value of this state-action pair is greater than the constraint threshold, we don’t need to enlarge the Qc-value, set $\alpha = 0$ works. Otherwise, if $Q_{c}^*(s, a) \leq l$, the choice of $\alpha$ that guarantee $Q_{c}^*(s, a) \geq l$ for $a \in A_c$, is then given by:
\[ \alpha \geq 2 \left( l - Q_{c}^*(s, a) \right) \left( I - \gamma P^\pi \right) \left[ \frac{\pi_\beta(a|s)}{\mu(a|s)} \right] (s, a) \]
\[ \Rightarrow \alpha \geq 2 \epsilon \cdot \max_{s,a} \left( l - Q_{c}^*(s, a) \right) \left( I - \gamma P^\pi \right) (s, a) \] (7)
Note that $(I - \gamma P^\pi)$ is a matrix (the inverse of the state occupancy matrix (Sutton, Barto et al. 1998)) with all non-negative entries, and (7) holds because of the definition of out-of-distribution action set. In all, choose $\alpha \geq \max \{ 2 \epsilon \max_{s,a} \left( l - Q_{c}^*(s, a) \right) \left( I - \gamma P^\pi \right) (s, a), 0 \}$ will satisfy $\forall s \in B, a \in A_c, \hat{Q}_{c}^*(s, a) \geq l$.

Now we show the error bound of the difference between the value obtained by CPQ and the value of the optimal safe policy $\pi^*$ on the dataset.

**Theorem 2.** Let $\| \hat{Q}_{c}^k - T^\pi \hat{Q}_{c}^{k-1} \|_{\mu, \beta}$ be the squared approximation error of the Constraints Penalized Bellman operator $T^\pi$ at iteration $k$. Let $\| Q_{c}^k - T^\pi Q_{c}^{k-1} \|_{\mu, \beta}$ be the squared approximation error of the Bellman evaluation operator $T^\pi$ at iteration $k$. If these two errors are bounded by $\delta$, then $\forall s \in B, a \in A_c, \hat{Q}_{c}^*(s, a) \geq l$.

1) $\lim_{k \to \infty} \hat{V}_{c}^{k} \leq l$; 2) $\lim_{k \to \infty} \left| V_{c}^{*} - \hat{V}_{c}^{k} \right| \leq \frac{4\gamma}{(1 - \gamma)^3} G(\epsilon) \sqrt{\delta}$
where $G(\epsilon) = \sqrt{(1 - \gamma)/\gamma} + \sqrt{\epsilon}/g(\epsilon)$ and define $g(\epsilon) := \min_{\mu^\beta(s)} \{ \mu^\beta(s) \}$, $g(\epsilon)$ captures the minimum discounted visitation probability of states under behaviour policy.

Proof. For 1), it can be easily derived from (4) that when $k \to \infty$, $\forall s \in B$, we have $\hat{V}_{c}(s) = \mathbb{E}_{a \sim P^\pi(s)} \left[ Q_{c}^*(s, a) \right] \leq l$. For 2), we give a proof sketch here, detailed proof can be found in Appendix A. The proof sketch goes as follow, we first convert the performance difference between $\pi^*$ and $\pi_\alpha$ to a value function gap that is filtered by the indicator function $1 \left( \hat{Q}_{c}(s', a') \leq l \right)$ (for simplicity, we denote it as $P_c$ below):
\[ V_{c}^{*} - \hat{V}_{c}^{k} \leq \frac{1}{\gamma} \sum_{s \in \mathcal{S}} \mathbb{E}_{a \sim h_{\pi_\alpha}(s)} \left[ P_{c}(s, a) \left( Q_{c}^*(s, a) - \hat{Q}_{c}^*(s, a) \right) \right] \]
where $\nu$ is any distribution over state-action space $\mathcal{S} \times \mathcal{A}$. We then prove $\left| P_{c}(Q_{c}^* - \hat{Q}_{c}^*) \right|_{\nu} = \mathbb{E}_{\nu} \left| P_{c}(Q_{c}^* - \hat{Q}_{c}^*) \right|_{\nu}$ can be bounded by $C \left( |Q_{c}^* - T^\pi Q_{c}^{k-1}|_{\mu, \beta} + |Q_{c}^* - T^\pi Q_{c}^*|_{\mu, \beta} \right)$.

Summary We show that we can enlarge Qc-values of OOD actions to be greater than $l$ by adjusting $\alpha$ in Theorem 1. We also show the performance guarantee in Theorem 2. Note that we can vary $\epsilon$ to make $G(\epsilon)$ as small as possible by adjusting the parameter $d$ in Algorithm 1, which is the only hyperparameter that needs to be tuned. This result guarantees that, upon the termination of Algorithm 1, the true performance of the main objective can be close to that of the optimal safe policy. At the same time, the safe constraint will be satisfied, assuming sufficiently large $k$.

**Experiments**

**Settings**

We conducted experiments on three Mujoco tasks: Hopper-v2, HalfCheetah-v2 and Walker2d-v2. These tasks imitate scenarios encountered by robots in real life. The robot is composed of multiple joints, at each step the agent selects the amount of torque to apply to each joint. In the experiments, we aim to prolong the motor life of different robots, while still enabling them to perform tasks. To do so, the motors of robots need to be constrained from using high torque values. This is accomplished by defining the constraint $C$ as the discounted cumulative torque that the agent has applied to each joint, and per-state penalty $c(s, a)$ is the amount of torque the agent decides to apply at each step.

For each environment, we collect data using a safe policy which has low rewards with safety constraints satisfied and an unsafe policy which has high reward but violates safety constraints. The unsafe policy was trained by PPO (Schulman et al. 2017) until convergence to the returns mentioned in Figure 1 and the safe policy was trained by CPO (Achiam et al. 2017) using the constraint threshold $l = 30$. The dataset is a mixture of 50% transitions collected by the safe policy and 50% collected by the unsafe policy. Mixture datasets are...
of particular interest, as it covers many practical use-cases where agents act safely in most cases but have some unsafe attempts for more benefits. Each dataset contains 2e6 samples. We used the same dataset for evaluating different algorithms to maintain uniformity across results.

Each agent is trained for 0.5 million steps and evaluated on 10 evaluation episodes (which were separate from the train distribution) after every 5000 iterations, we use the average score and the variance for the plots.

**Baselines**

We compare CPQ with the following baselines:

- **CBPL**: CBPL (Le, Voloshin, and Yue 2019), which learns safe policies by applying FQE and FQI, was originally designed for discrete control problems, we extend it to continuous cases by using continuous FQI (Antos, Szepesvári, and Munos 2008).

- **BCQ-Lagrangian**: As BCQ (Fujimoto, Meger, andPrecup 2019) was not designed for safe offline RL, we combine BCQ with the Lagrangian approach, which uses adaptive penalty coefficients to enforce constraints, to obtain BCQ-Lagrangian.

- **BEAR-Lagrangian**: Analogous to BCQ-Lagrangian, but to use another state-of-the-art offline RL method BEAR (Kumar et al. 2019).

- **BC-Safe**: As mentioned in Section 3.2, we also include a Behavior Cloning baseline, using only data generated from the safe policy. This serves to measure whether each method actually performs effective RL, or simply copies the data.

**Comparative Evaluations**

It is shown in Figure 1 that CPQ achieves higher reward while still satisfying safety constraints, compared to two naïve approaches (BCQ-L and BEAR-L). Naïve approaches achieve sub-optimal performance due to the reasons discussed in Section 3.2. For example, BEAR-L has difficulty to learn the balance between two Lagrangian multipliers, $\lambda_1$ for the safety constraint and $\lambda_2$ for the divergence constraint, this two multiplier raise their value by turns to try to satisfy either of the two constraints, making the effect of Qr diluted. BCQ-L performs better than BEAR-L, but still suffers from zig-zag learning curves due to the similar reason. CBPL diverges and fails to learn good policies in all three environments, due to large value estimation errors caused by OOD actions. It can also be observed that the constraint values of CPQ are sometimes lower than the threshold, due to the reason that $\nu$ sometimes falsely chooses in-distribution actions, making the Qr-values of these actions erroneously large. This suggests that the performance of CPQ may be further enhanced by applying more advanced OOD detection techniques to construct $\nu$, we leave it for future work.

**Sensitivity to Constraint Limit $l$**

The results discussed in the previous section suggest that CPQ outperforms other baselines on several challenging tasks. We’re now interested in the sensitivity of CPQ to different constraint limit $l$. It can be seen in Figure 2 that CPQ is robust to different constraint limits. This means that we can do counterfactual policy learning\(^2\) (García and Fernández 2015), i.e., adjust $l$ post-hoc to derive policies with different safety requirements. Note that imitation-based methods (e.g., BC-Safe) can only satisfy the original constraint limit $l$.

**Conclusions and Future Work**

We present a novel safe offline RL algorithm, CPQ, the first continuous control RL algorithm capable of learning from mixed offline data under constraints. Through theoretical analysis as well as systematic experimental results, we show that CPQ achieves better performance across a variety of tasks, comparing to several baselines. One future work is to use more advanced OOD detection techniques (e.g., using

\(^2\)In online RL under constraint, the agent needs to “re-sample-and-learn” from scratch when the constraint limit is modified.
energy scores (Liu et al. 2020), to further enhance CPQ’s performance. Another future work is developing new algorithms to tackle offline RL under hard constraints. We hope our work can shed light on safe offline RL, where one could train RL algorithms offline, and provides reliable policies for safe and high quality control in real-world tasks.

Acknowledgments

A preliminary version of this work was accepted as a spotlight paper in RL4RealLife workshop at ICML 2021. This work was supported by the National Key R&D Program of China (2019YFB2103201) and the National Science Foundation of China (No. 61502375).

References

Achiam, J.; Held, D.; Tamar, A.; and Abbeel, P. 2017. Constrained policy optimization. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, 22–31. JMLR. org.

Altmann, E. 1999. Constrained Markov decision processes, volume 7. CRC Press.

Antos, A.; Szepesvári, C.; and Munos, R. 2008. Fitted Q-iteration in continuous action-space MDPs. In Advances in neural information processing systems, 9–16.

Argenson, A.; and Dulac-Arnold, G. 2021. Model-Based Offline Planning. In International Conference on Learning Representations.

Chow, Y.; Ghavamzadeh, M.; Janson, L.; and Pavone, M. 2017. Risk-constrained reinforcement learning with percentile risk criteria. The Journal of Machine Learning Research, 18(1): 6070–6120.

Dalal, G.; Dvijotham, K.; Vecerik, M.; Hester, T.; Paduraru, C.; and Tassa, Y. 2018. Safe exploration in continuous action spaces. arXiv preprint arXiv:1801.08757.

Fujimoto, S.; Hoof, H.; and Meier, D. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In International Conference on Machine Learning, 1587–1596.

Fujimoto, S.; Meier, D.; and Precup, D. 2019. Off-policy deep reinforcement learning without exploration. In International Conference on Machine Learning, 2052–2062. PMLR.

Garcia, J.; and Fernández, F. 2015. A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 16(1): 1437–1480.

Ha, S.; Xu, P.; Tan, Z.; Levine, S.; and Tan, J. 2020. Learning to Walk in the Real World with Minimal Human Effort. arXiv preprint arXiv:2002.08550.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C. P.; Glorot, X.; Botvinick, M. M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In International Conference on Learning Representations.

Kidambi, R.; Rajeswaran, A.; Netrapalli, P.; and Joachims, T. 2020. MOReL: Model-Based Offline Reinforcement Learning. In Neural Information Processing Systems (NeurIPS).
Tessler, C.; Mankowitz, D. J.; and Mannor, S. 2018. Reward Constrained Policy Optimization. In International Conference on Learning Representations.

Watkins, C. J.; and Dayan, P. 1992. Q-learning. Machine learning, 8(3-4): 279–292.

Wu, Y.; Tucker, G.; and Nachum, O. 2019. Behavior Regularized Offline Reinforcement Learning. arXiv preprint arXiv:1911.11361.

Yu, T.; Thomas, G.; Yu, L.; Ermon, S.; Zou, J.; Levine, S.; Finn, C.; and Ma, T. 2020. MOPO: Model-based Offline Policy Optimization. In Neural Information Processing Systems (NeurIPS).

Zhan, X.; Xu, H.; Zhang, Y.; Huo, Y.; Zhu, X.; Yin, H.; and Zheng, Y. 2021. DeepThermal: Combustion Optimization for Thermal Power Generating Units Using Offline Reinforcement Learning. arXiv preprint arXiv:2102.11492.

Zhan, X.; Zhu, X.; and Xu, H. 2021. Model-Based Offline Planning with Trajectory Pruning. arXiv preprint arXiv:2105.07351.