CONFORMAL ACTOR-CRITIC: DISTRIBUTION-FREE UNCERTAINTY QUANTIFICATION FOR OFFLINE RL

Alok Shah[1*], **Nikhil Kumar**[1*], **Khush Gupta**[1*]
[1]University of Pennsylvania
{alokshah, kumarnik}@sas.upenn.edu
khushg@wharton.upenn.edu
* *denotes equal contribution*

ABSTRACT

Offline Reinforcement Learning (RL) aims to learn policies from fixed datasets without active environmental interaction, but often suffers from overestimated Q-values on out-of-distribution (OOD) state-action pairs. In this paper, we propose *Conformal SAC*, a lightweight and theoretically grounded approach that integrates *Conformal Prediction*, a distribution-free uncertainty quantification method, into the Soft Actor-Critic (SAC) model for offline RL. Conformal SAC constructs finite-sample prediction intervals around Q-value estimates, using their width to adaptively down-weight uncertain actions during policy learning. Unlike ensemble or bootstrap-based uncertainty estimates, conformal prediction is computationally efficient and enjoys formal coverage guarantees without drastic changes to the underlying RL architecture. We provide theoretical support for both marginal and group-conditional coverage, and demonstrate improved stability and robustness across benchmark control tasks. In particular, we show how conformal intervals mitigate erratic behavior in bang-bang control problems by explicitly accounting for uncertainty in the learned value function. Empirically, our approach achieves competitive performance with strong model-free offline RL baselines, highlighting how tools from statistical inference and causal reasoning can promote safer, more reliable offline policy learning with minimal overhead.

1 INTRODUCTION

Reinforcement Learning (RL) has demonstrated remarkable success across a wide range of sequential decision-making tasks—from robotic control to video games to large language model alignment (Sutton & Barto, 2018; Ziegler et al., 2019). However, its real-world deployment is often hindered by the impracticality, cost, or danger of direct environment interaction (Dulac-Arnold et al., 2019). In such settings, Offline RL provides an appealing alternative: policies are trained from pre-collected datasets without further environment access (Levine et al., 2020). Yet, this paradigm introduces new challenges, particularly the problem of distributional shift. When a learned policy encounters out-of-distribution (OOD) state, action pairs, the Q-function rapidly increases (due to the $\max$ operator within the Bellman update), producing over-optimistic value estimates at the expense of generalization(Fujimoto et al., 2019; Kumar et al., 2020a).

To address this, several approaches have been proposed. Behavior-constrained algorithms like TD3+BC (Fujimoto & Gu, 2021), AWAC (Nair et al., 2020), and IQL (Kostrikov et al., 2021) guide the policy to stay close to the data, often at the cost of limited exploration. Other methods inject pessimism directly into value estimates to avoid unreliable OOD actions. Conservative Q-Learning (CQL) (Kumar et al., 2020a), for instance, penalizes Q-values on unseen actions, while ensemble-based methods like EDAC (An et al., 2021) and dropout-based techniques like UWAC (Wu et al., 2021) use heuristic uncertainty estimates to regularize training. While effective, these approaches either lack statistical coverage guarantees or require significant computational overhead.

We propose Conformal SAC, a simple yet theoretically grounded alternative that integrates conformal prediction into model-free actor-critic methods. Using split-conformal calibration (Lei et al., 2017; Romano et al., 2019; Angelopoulos & Bates, 2021), we construct finite-sample prediction in-

tervals around Q-values, enabling the policy to avoid uncertain actions and safely exploit confident ones. Our approach retains the architectural simplicity of standard actor-critic algorithms and does not require ensembles, behavior cloning, or explicit dynamics modeling. Additionally, we extend our method to achieve group-conditional coverage, enabling localized uncertainty calibration over subsets of the state-action space.

We establish theoretical guarantees for both marginal and group-based coverage of Q-values and demonstrate empirically that our approach improves robustness, reduces overestimation, and enhances policy stability across benchmark offline RL tasks. In particular, we validate the method on bang-bang control problems—where optimal actions lie on the boundary of the action space—highlighting how conformal prediction provides safe, data-driven conservatism.

2 Related Works

Offline Reinforcement Learning has made significant progress in addressing distributional shift and extrapolation error, which arise when policies evaluate or act on out-of-distribution (OOD) state-action pairs. Model-free approaches like Conservative Q-Learning (CQL) Kumar et al. (2020a) tackle this by penalizing Q-values for OOD actions, encouraging conservative estimates. However, CQL's fixed penalty can limit expressiveness and lead to underutilization of good but rare actions. Other approaches trade off conservatism and behavior cloning: TD3+BC Fujimoto & Gu (2021) adds a weighted BC loss to TD3, while AWAC Nair et al. (2020) and IQL Kostrikov et al. (2021) leverage advantage-weighted behavior cloning to selectively reinforce strong actions from the dataset. Ensemble-based methods like EDAC An et al. (2021) improve epistemic uncertainty estimates via Q-network diversification, while dropout-based approaches such as UWAC Wu et al. (2021) offer lower overhead approximations. These methods aim to mitigate overestimation by using heuristics or pessimism to implicitly model uncertainty. However, they lack formal coverage guarantees and often require expensive ensemble training or behavior regularization. Model-based methods like MOPO Yu et al. (2020) and MOReL Kidambi et al. (2020) enforce conservatism by generating pessimistic rollouts, but suffer from model bias and compounding error, especially on high-dimensional or complex tasks.

Conformal Prediction provides a complementary approach by offering distribution-free uncertainty quantification with finite-sample coverage guarantees. Originally developed for supervised settings Vovk et al. (2018), conformal prediction techniques construct predictive intervals that contain the true output with a user-specified probability, without strong assumptions on the data distribution. Split conformal prediction Lei et al. (2017); Romano et al. (2019); Angelopoulos & Bates (2021) has emerged as a practical variant, requiring only a calibration set to determine quantile thresholds. In RL, conformal prediction has been explored for safe policy improvement Petrik et al. (2016), policy evaluation Thomas et al. (2015), and in some cases for distributional value estimation Dabney et al. (2018). However, these methods assume access to on-policy rollouts, focus on model-based regimes, or impose strong structural assumptions. Few have explored the use of conformal prediction directly within model-free actor-critic architectures, where value function errors can propagate and amplify in unexpected ways.

Our approach builds on these ideas by integrating conformal prediction into model-free offline RL, offering an efficient and principled method for quantifying Q-function uncertainty. We apply split conformal calibration to the critic's residuals, constructing prediction intervals that guide policy updates via uncertainty-aware penalization. Unlike CQL's fixed conservatism or ensemble/dropout-based heuristics, our method uses a single Q-network and provides finite-sample statistical coverage for Q-value estimates. Moreover, we extend this to group-conditional coverage via GroupSplitConformal, enabling localized uncertainty calibration across state-action subspaces. As such, Conformal Actor-Critic emerges as a more adaptive, optimistic alternative CQL armed with explicit uncertainty-quanitification.

3 PRELIMINARIES

3.1 OFFLINE REINFORCEMENT LEARNING

Offline Reinforcement Learning (Offline RL) addresses scenarios where an agent learns from a fixed dataset without interacting with the environment.

3.1.1 MARKOV DECISION PROCESS (MDP)

As in standard RL literature, an MDP is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ defines the state space, $\mathcal{A}$ defines the action space, $P(s'|s,a)$ denotes the transition probability function, $|R(s,a)| \leq B_R$ represents the reward function R whose values are bounded by B_R, and $\gamma \in [0,1]$ denotes the discount factor. The goal is to find a policy $\pi(a|s)$ that maximizes the expected cumulative reward:

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)\right]. \tag{1}$$

where $\tau = (s_0, a_0, s_1, a_1, \dots)$ denotes a trajectory.

We assume access to a dataset $D = (s, a, r, s')$ of tuples generated from trajectories sampled from a behavior policy π_β which typically correspond to some level of skill

3.1.2 BELLMAN EQUATION

The Bellman equation describes how to compute optimal policies in an MDP. Since the data is Markovian, it suffices to use the principle of dynamic programming, making locally optimal choices at each time step yields a globally optimal solution.

Formally, the action-value function $Q^\pi(s,a)$ measures the expected return when starting from state s, taking action a, and following policy π thereafter:

$$Q^\pi(s,a) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t) \mid s_0 = s, a_0 = a\right] \tag{2}$$

We denote the value function V induced by the policy π as:

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi(s,a)\right] \tag{3}$$

The optimal action-value $Q^*(s,a)$ respects **Bellman Optimality**:

$$Q^*(s,a) = \mathbb{E}_{\tau\sim P}\left[R(s,a) + \gamma \max_{a'} Q^*(s',a')\right] \tag{4}$$

This expresses the expected cumulative reward from taking action a in state s, receiving immediate reward $R(s,a)$, and continuing optimally from the next state s'.

3.1.3 POLICY AND Q-FUNCTION UPDATES UNDER ACTOR CRITIC

Q-learning methods parameterize the action-value function $Q_\theta(s,a)$ and update it through the Bellman backup operator:

$$Q_\theta(s,a) \leftarrow R(s,a) + \gamma \max_{a'} Q_\theta(s',a') \tag{5}$$

Actor-Critic methods extend this by maintaining a separate policy $\pi_\phi(a|s)$. The policy π_ϕ is updated to maximize expected Q-values, while the Q-function is updated using samples from the dataset:

Policy Evaluation (Q-function Update):

$$Q_\theta^{k+1} \leftarrow \arg\min_Q \mathbb{E}_{(s,a,r,s')\sim D}\left[\left(R(s,a)+\gamma\mathbb{E}_{a'\sim\pi_\phi(a'|s')}[Q_\theta^k(s',a')]-Q(s,a)\right)^2\right] \tag{6}$$

Note that equation 6 simply minimizes the temporal difference error, that is, the MSE between the Bellman backup target and the predicted Q-value.

Policy Improvement (Policy Update):

$$\pi_\phi^{k+1} \leftarrow \arg\max_\pi \mathbb{E}_{s\sim D}\left[\mathbb{E}_{a\sim\pi(a|s)}[Q_\theta^{k+1}(s,a)]\right] \tag{7}$$

3.1.4 Extrapolation Error

Given that D is finite, it follows quite naturally that training suffers from distribution shifs in the actions. The Bellman backup requires evaluating $Q_\theta(s',a')$ for actions $a'\sim\pi_\phi(a'|s')$, which may not be present in the dataset. Since the Q-function is trained on in-distribution actions only, it can overestimate the value of out-of-distribution (OOD) actions due to the max operator:

$$Q_\theta(s,a) = R(s,a)+\gamma\max_{a'} Q_\theta(s',a') \tag{8}$$

In the online setting, one can correct for this by taking the action, observing the resulting lower value, and taking the correct gradient step. However, the nature of the Offline RL setting with a priori trajectory data does not afford this flexibility.

3.1.5 CQL

For the purposes of theoretical analysis, we'll focus on analysis of Conservative Q-Learning (Kumar et al., 2020a), which addresses extrapolation error by penalizing Q-values for OOD actions. The CQL objective modifies the standard Q-function updates with:

$$\begin{aligned} Q_\theta^{k+1} = \arg\min_Q\ & \alpha\mathbb{E}_{s\sim D,a\sim\mu(a|s)}[Q(s,a)] \\ & + \frac{1}{2}\mathbb{E}_{(s,a,r,s')\sim D}\left[Q(s,a)-\left(R(s,a)+\gamma\max_{a'}Q_\theta^k(s',a')\right)\right]^2 \end{aligned} \tag{9}$$

The parameter α controls OOD penalty magnitude and $\mu(a\mid s)$ denotes a sampling distribution over actions, chosen to include OOD actions.

To avoid penalizing in-distribution actions, CQL introduces a correction term based on the behavior policy $\pi_\beta(a\mid s)$, giving:

$$\begin{aligned} Q_\theta^{k+1} = \arg\min_Q\ & \alpha\left(\mathbb{E}_{s\sim D,a\sim\mu(a|s)}[Q(s,a)]-\mathbb{E}_{s\sim D,a\sim\pi_\beta(a|s)}[Q(s,a)]\right) \\ & + \frac{1}{2}\mathbb{E}_{(s,a,r,s')\sim D}\left[Q(s,a)-\left(r+\gamma\max_{a'}Q_\theta^k(s',a')\right)\right]^2 \end{aligned} \tag{10}$$

Here, $\mathbb{E}_{s\sim D,a\sim\pi_\beta(a|s)}[Q(s,a)]$ denotes the expected Q-value of actions likely based on the behavior policy. The subtraction term offsets the penalty on in-distribution actions, making the learned Q-function more conservative only for uncertain or OOD actions.

3.2 Conformal Prediction

Conformal prediction constructs prediction intervals for a model's outputs. Given a calibration dataset D_{cal} and a predictor f, split conformal prediction computes a confidence interval $C(s,a)$:

$$\hat{C}(s,a) = [f(s_i,a_i)-q_\alpha, f(s_i,a_i)+q_\alpha], \tag{11}$$

where q_α is the $(1-\alpha)$ quantile of nonconformity scores $\alpha_i = |f(s_i,a_i)-y_i|$ over D_{cal}.

Using split conformal prediction in our setting, we construct a conformal prediction interval for an unseen state-action pair (s_{m+1},a_{m+1}) as follows:

$$\hat{C}(s_{m+1},a_{m+1}) = [f(s_{m+1},a_{m+1})-q_\alpha, f(s_{m+1},a_{m+1})+q_\alpha] \tag{12}$$

4 CONFORMAL ACTOR-CRITIC

Conformal Actor-Critic introduces statistical coverage guarantees by integrating conformal prediction into an actor-critic framework, while addressing extrapolation error. Our algorithm integrates conformal intervals into the standard actor-critic procedure.

We start with an offline dataset and designed calibration subset. In each training iteration, we first compute the $(1–\alpha)$-quantile of our Q-value estimation errors, which serves as a measure of the model's uncertainty. The subsequent actor and critic updates performed on a sampled minibatch are relatively standard; we update the Q-network is updated by minimizing MSE across the minibatch, and the policy network is updated by maximizing the expected Q-values while incorporating a regularization term proportional to the uncertainty measure.

The algorithm pseudo-code is presented below.

Algorithm 1 Conformal Soft Actor-Critic (Conformal SAC)

Input: Offline dataset $D = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{N}$, calibration set $D_{\text{calib}} \subset D$, confidence level $1 - \alpha$, smoothing factor τ
Initialize policy π_ϕ, Q-networks $Q_{\theta_1}, Q_{\theta_2}$, value network V_ψ, target network $V_{\psi'} \leftarrow V_\psi$
Initialize conformal threshold $q_\alpha \leftarrow 0$
for each training iteration **do**
  Sample batch $B = \{(s, a, r, s', d)\} \subset D$
  Critic (Q-function) Update:
  Estimate targets:

$$y = r + \gamma(1 - d) \cdot V_{\psi'}(s')$$

  Compute MSE loss:

$$\mathcal{L}_{Q_i} = \mathbb{E}\left[\left(Q_{\theta_i}(s, a) - y\right)^2\right]$$

  Update $Q_{\theta_1}, Q_{\theta_2}$ using gradient descent
  Policy Update:
  Sample $a \sim \pi_\phi(\cdot|s)$ and compute:

$$\mathcal{L}_\pi = \mathbb{E}_s\left[\eta \log \pi_\phi(a|s) - \min(Q_{\theta_1}(s, a), Q_{\theta_2}(s, a))\right]$$

  Update policy π_ϕ
  Value Function Update:
  Estimate $Q_{\text{min}} = \min(Q_{\theta_1}(s, a), Q_{\theta_2}(s, a))$
  Estimate entropy-regularized target:

$$y_V = Q_{\text{min}} - \eta \log \pi_\phi(a|s)$$

  Apply conformal penalty:

$$y_V' = y_V \cdot (1 - \tilde{q}_\alpha)$$

  where $\tilde{q}_\alpha = \frac{q_\alpha}{\texttt{std}(y_V) + \epsilon}$
  Update V_ψ by minimizing:

$$\mathcal{L}_V = \mathbb{E}\left[\left(V_\psi(s) - y_V'\right)^2\right]$$

  Target Network Update: $\psi' \leftarrow \tau\psi + (1 - \tau)\psi'$
  Calibrate Conformal Threshold:
  Sample $(s, a, r, s') \in D_{\text{calib}}$ and compute residuals:

$$\alpha_i = |Q_{\theta_1}(s, a) - (r + \gamma V_{\psi'}(s'))|$$

  Set $q_\alpha \leftarrow \text{Quantile}(\{\alpha_i\}, 1 - \alpha)$
end for

We also consider a more adaptive algorithm, which calibrates the choices of quantiles based on groups. Similar to canonical group constructions, we can consider groups here as similar clusters of state-action pairs. Let $G \subseteq 2^{\mathcal{S} \times \mathcal{A}}$ be some collection of groups. We assume such groups $G_j \in \mathcal{G}$ to

be well-defined, i.e. sufficiently large and incomparable (meaning no group is a subset of another), ensuring that these groups generally cover the state-action space.

Since we want coverage over each group, we choose our conformal interval length as a function of the group/state-action pair. Therefore, we introduce another function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ that determines the threshold for that pair, specifying the interval construction. Define $g(s, a)$ as:

$$g(s,a) = \sum_{j=1}^{J} \lambda_j \; g_j(s,a) \tag{13}$$

where $g_j(s, a)$ is an indicator variable specifying if the state-action pair belongs to group j. We define the GroupSplitConformal algorithm below.

Algorithm 2 Group Offline Conformal Actor-Critic (GroupSplitConformal)

Input: Offline dataset $D = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{N}$, calibration set $D_{\text{cal}} \subseteq D$, groups $\{G_j\}_{j=1}^{J}$, regularization parameter η, confidence level $1 - \alpha$
Initialize Q-network Q_θ, policy network π_ϕ, and learning rates
for each training iteration **do**
    Sample a minibatch $B \subseteq D$
    Critic Update (Q-network):

$$\mathcal{L}_Q = \frac{1}{|B|} \sum_{(s,a,r,s') \in B} \left(r + \gamma \mathbb{E}_{a' \sim \pi_\phi(\cdot|s')} Q_\theta(s', a') - Q_\theta(s,a)\right)^2$$

    Update Q_θ by minimizing $\mathcal{L}_Q$
    Fit Group-Adaptive Thresholds:
    Compute nonconformity scores for D_{cal}: $\alpha_i = |Q_\theta(s_i, a_i) - y_i|$
    Let λ^* be the solution to the optimization problem:

$$\lambda^* = \arg\min_{\lambda} \mathbb{E}_{(s,a) \sim D_{\text{cal}}} \left[L_{1-\alpha}(g(s;\lambda), \alpha)\right] + \eta \|\lambda\|_1,$$

    where

$$g(s,a;\lambda) = \sum_{j=1}^{J} \lambda_j g_j(s,a),$$

    and $L_{1-\alpha}$ is a loss function ensuring quantile consistency within each group.
    Set $g(s, a) = g(s, a; \lambda^*)$
    Actor Update (Policy Network):
    Given $g(s, a)$, define intervals $\hat{C}(s, a) = [Q_\theta(s, a) - g(s, a), Q_\theta(s, a) + g(s, a)]$.
    Incorporate g into the actor loss:

$$\mathcal{L}_\pi = -\frac{1}{|B|} \sum_{s \in B} \mathbb{E}_{a \sim \pi_\phi(\cdot|s)}[Q_\theta(s,a)] + \lambda^* \mathbb{E}_{(s,a) \sim D_{\text{cal}}}[g(s,a)],$$

    Update π_ϕ by minimizing $\mathcal{L}_\pi$.
end for

4.1 THEOREMS

We present theoretical guarantees to establish that our conformal prediction algorithm provides finite-sample coverage guarantees and constructs prediction intervals that are effective for uncertainty quantification.

Assumption 1 (Consistent Q-value estimates) *The Q-value labels y_i are consistent estimates of $Q^\pi(s_i, a_i)$, i.e. $\mathbb{E}[|y_i - Q^\pi(s_i, a_i)|] \to 0$ as $|D_{cal}| \to \infty$*

[1]

[1] Standard in RL theory, with sufficient calibration data the estimation error should vanish. Note that without such consistency, any learning/Bellman updating will always have non-trivial errors.

Definition 1 (Nonconformity Score) *For a state-action pair (s, a), define the nonconformity score as:*

$$\alpha = |f(s, a) - y_i|$$

where $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the learned Q-value predictor.

We want to learn the associated Q-values for all state action pairs. For each (s_i, a_i) in our dataset D, we let y_i be the target Q-value for that state action pair. With access to such labels, optimizing over our Bellman equation is trivial.

Theorem 1 (Conformal Coverage) *Suppose that the calibration dataset $D_{cal} = (s, a, r, s')$ of size m is exchangeable. Consider the test point $(s_{m+1}, a_{m+1}, y_{m+1})$. Furthermore, let q_α be the $(1 - \alpha)(1 + \frac{1}{m})$-quantile for each score q_α Then, the conformal prediction interval $\hat{C}(s, a) = [f_\theta(s, a) - q_\alpha, f_\theta(s, a) + q_\alpha]$ from* **??** *satisfies the standard marginal coverage guarantee:*

$$P(y_{m+1} \in \hat{C}(s_{m+1}, a_{m+1})) \in [1 - \delta, 1 - \delta + \frac{1}{m + 1}] \tag{14}$$

Proof Sketch: Similar to standard conformal marginal coverage proofs; sort nonconformity scores in increasing order, and use uniformity of scores and exchangeable data points to show lower and upper bound. See more detailed proofs in Appendix A.

We now present a group-conditional coverage guarantee. Under GroupSplitConformal, we have the following conditional group guarantee.

Theorem 2 (Group-Conditional Coverage) *Let $\mathcal{G} = \{G_j\}_{j=1}^{J}$ be a collection of groups defined over $\mathcal{S} \times \mathcal{A}$. Assume that each group G_j is sufficiently large and incomparable (i.e., no group is a subset of another). Applying the GroupSplitConformal algorithm, the prediction intervals satisfy:*

$$1 - \alpha - \epsilon_j \leq \mathbb{P}\left(y \in \hat{C}(s, a) \mid (s, a) \in G_j\right) \leq 1 - \alpha + \epsilon_j,$$

where $\epsilon_j = O(\sqrt{\frac{\alpha}{\mu(G_j)}})$ and $\mu(G_j)$ denotes the proportion of the calibration data in group G_j.

Proof Sketch: Follows from Proof of Theorem 35 in Roth (2024). By assigning adaptive thresholds $g(s, a)$ that depend on the group membership of (s,a), the resulting intervals achieve approximately valid conditional coverage in each group. The error term ϵ vanishes as the calibration data size within each group grows.

We now present a few theorems relating Conformal Actor-Critic to related literature, particularly the CQL framework. As mentioned above, CQL adds a regularization term to the Bellman update. We generally show that Conformal Actor-Critic is more optimistic than CQL, but also remaining conservative.

In particular, we show that if the policy uses the lower end of the prediction intervals to make decisions, the resulting policy performs close to the behavior policy. Define the **lower confidence bound** as $\mathcal{L}(s, a) := f(s, a) - q_\alpha$ where f and q_α defined analogously to above.

Theorem 3 (Conformal Actor-Critic is Conservative) *Assume that labels y_i are consistent, and Theorem 1 holds. Define the policy $\pi_L(a|s) := \arg\max_{a'} \mathcal{L}(s, a')$. Then, the value function resulting from π_L approaches closer to the value function from the behavior policy as the dataset size grows. Formally, there exists a function $\epsilon(m, \alpha)$ such that with probability $\geq 1 - \alpha$,*

$$V_{\pi_L}(s) \geq V_{\pi_\beta}(s) - \epsilon(m, \alpha) \quad \forall s \in \mathcal{S}$$

where $\epsilon(m, \alpha)$ is the error gap, and $\epsilon(m, \alpha) \to 0$ as $m \to \infty$.

Proof Sketch: $\mathcal{L}(s, a)$ is a lower confidence bound for $Q^\pi(s, a)$ w.h.p. and as actions are selected based on the adaptive policy, the difference only stems from dataset size.

Theorem 4 (Conformal Actor-Critic is More Optimistic than CQL) *Consider standard Conservative Q-Learning (CQL) which learns a Q-function $Q_{CQL}(s, a)$ with a uniform regularization term. Suppose as $|D| \to \infty$, the learned predictor $f_\theta(s, a)$ converges to $Q^\pi(s, a)$ and $q_\alpha \to 0$. Then, with probability $\geq 1 - \alpha$,*

$$\mathcal{L}(s, a) = f_\theta(s, a) - q_\alpha \geq Q_{CQL}(s, a) \quad \forall (s, a) \in \textit{supp}(\pi_\beta).$$

Proof Sketch: CQL ensures no worse than behavior policy performance if Q-values are underestimated. Here, the conformal interval lower endpoint $f(s,a) - q_\alpha$ acts like a lower bound dependent on the calibration dataset. Since it contains the true Q-value with high probability, selecting actions that maximize this lower bound avoids actions that would drastically reduce return. As our conformal intervals become tighter with more data, the conformal lower bound can exceed the uniformly lowered Q-values from CQL, allowing more optimism where justified by the data.

We note that in the CQL framework, a uniform regularization penalty is added to avoid overestimation. With Conformal Actor-Critic, the intervals are constructed such that the regularization-equivalent term q_α – the length of the interval – shrinks as the dataset size grows. In our context, the computed quantiles q_α serve a similar purpose to the regularization term in CQL; however, since these quantiles converge to 0 as the dataset size increases, we are able to show more optimistic estimates due to conformal prediction's data dependent approach.

4.2 BANG-BANG CONTROLLERS

Bang-bang controllers are a class of control systems that operate by switching between extreme states or actions, rather than continuously modulating control inputs. The most well-known application of bang-bang controllers is in thermostats, where the whole system turns on if the current temperature is above the desired threshold, and remains off otherwise. In classical control theory, time-optimal control problems for linear systems with bounded inputs often result in bang-bang solutions. Specifically, consider a linear system governed by

$$\dot{x}(t) = Ax(t) + Bu(t),$$

where $u(t) \in \mathcal{U} = \{-1, 1\}^m$. Under the analysis of Liberzon (2011), the Pontryagin Maximum Principle ensures that the optimal control $u^*(t)$ switches between its extreme values of 1 and -1, and the number of sign switches is finite. This means that the optimal solution lies on the boundary of the convex hull.

Although such controllers can be simple in construction, they can exhibit abrupt behavior that may lead to instability or unsafe operations. To address this, we apply Conformal Actor-Critic to bang-bang controllers, showing that the constructed prediction intervals can result in less volatile and unstable behavior. Since action choices all lie on the boundary of the convex hull of the action space, Conformal Actor-Critic allows us to quantify uncertainty in Q-values of certain state-action pairs and thus have a more stable system.

Mathematical Framework:

Consider a finite state space $\mathcal{S}$ and a discrete action space $\mathcal{A} = \{-a_{\max}, a_{\max}\}^d$, where $a_{\max} > 0$. Let the transition probability function $P(s'|s,a)$ and reward function $R(s,a)$ be defined as in Section 2.1.1. The policy function $\pi_\theta(a|s)$ is a bang-bang policy that selects actions from the corners of the d-dimensional action space hypercube.

In our Conformal Actor-Critic algorithms, after updating the Q-network to minimize the Bellman error, we compute conformal prediction intervals

$$\hat{C}(s,a) = [Q_\theta(s,a) - q_\alpha,\ Q_\theta(s,a) + q_\alpha],$$

using the calibration set D_{cal}, where q_α is the quantile computed from the calibration data. Then, during the policy update, the actor incorporates q_α to adjust the policy towards actions with lower uncertainty. The actor update which minimizes $\mathcal{L}_\pi$ in Algorithm **??** is particularly relevant. Note that by Theorem 1, conformal coverage holds under our discrete action space, and Theorem 2 and Algorithm 2 can be used to extend group guarantees. In particular, the state-action pairs can be partitioned into groups corresponding to a certain operating regime, i.e. a set of safe states $S_{\text{safe}} \subseteq \mathcal{S}$.

We now show that the estimated quantiles from the data are close to the actual quantiles. This follows from a straightforward application of the Dvoretzky–Kiefer–Wolfowitz (DKW) Inequality, which provides uniform convergence guarantees for the empirical distribution function.

To find q_α, we use the empirical CDF of nonconformity scores:

$$\hat{F}_\alpha(c) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\left[\alpha_i \leq c\right].$$

Note that the true CDF under the data distribution is:

$$F_\alpha(c) = \mathbb{P}(\alpha \leq c).$$

Claim 1 (Quantile Consistency) *Let $\{\alpha_i\}_{i=1}^m$ be i.i.d. samples from a distribution with CDF F_α. Define the true $(1-\alpha)$-quantile as*

$$q_\alpha := \inf\{c : F_\alpha(c) \geq 1-\alpha\},$$

and the empirical $(1-\alpha)$-quantile as

$$\hat{q}_\alpha := \inf\{c : \hat{F}_\alpha(c) \geq 1-\alpha\}.$$

Then, for any $\delta > 0$, with probability at least $1-\delta$,

$$|\hat{q}_\alpha - q_\alpha| \leq \epsilon,$$

where ϵ is chosen so that

$$F_\alpha(q_\alpha + \epsilon) - F_\alpha(q_\alpha - \epsilon) \leq 2t,$$

and $t = \sqrt{\frac{\ln(2/\delta)}{2m}}$.

Proof Sketch: Follows from the DKW Inequality and monotonicity of F_α.

Claim 2 (Uniform Convergence of Q-values) *Consider Q-value estimation errors defined as:*

$$e_i := f_\theta(s_i, a_i) - Q_\pi(s_i, a_i),$$

where $\{e_i\}_{i=1}^m$ are i.i.d. and bounded by $M > 0$. Let $\mathcal{S}$ be a finite state space and $\mathcal{A} = \{-a_{\max}, a_{\max}\}^d$ the discrete bang-bang action space. Then, for any $\delta > 0$, with probability at least $1-\delta$:

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |f_\theta(s,a) - Q_\pi(s,a)| \leq M\sqrt{\frac{2\ln\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{m}}.$$

Proof: Construct bounded normalized errors and use Hoeffding's Inequality with a relevant choice of t. Further detailed proofs for Claims 1 and 2 are provided in Appendix A.

By combining these theoretical results, we obtain a high-probability guarantee that the constructed conformal intervals will contain the true Q-values for every state-action pair. Moreover, when actions differ only slightly along one dimension, these intervals help ensure that action switches are made only when the system is sufficiently confident. This leads to policies that are both more stable and safer, as they mitigate abrupt and potentially hazardous control changes.

5 Experiments

5.1 Setup

We evaluate Conformal Actor-Critic on the D4RL benchmark suite (Fu et al., 2021), focusing on continuous control environments, which present distinct challenges in terms of reward structure and dynamics, offering a broad view of generalization and uncertainty estimation. We compare against standard offline RL baselines, including SAC (Haarnoja et al., 2018) and CQL (Kumar et al., 2020b)

Our implementation of Conformal Q-Learning uses a single Q-network and introduces uncertainty via split-conformal calibration with a held-out calibration set (10% of the offline dataset). Additional implementation details can be found in A.2

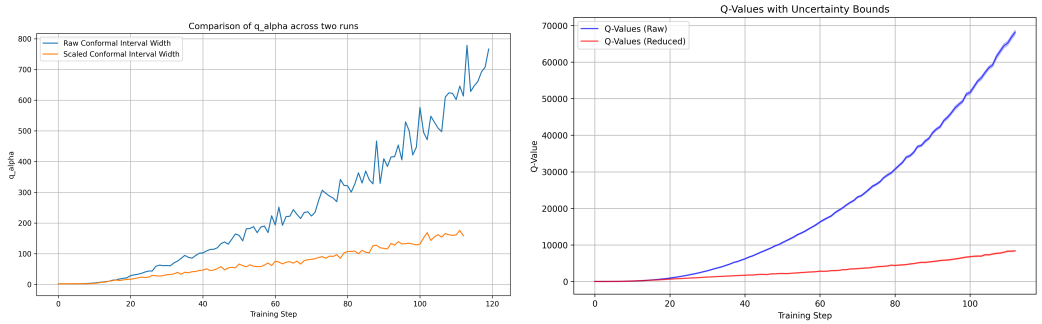
5.2 Main Results: Performance Across Benchmarks

For each method, we run 4 random seeds and report the mean score:

We observe a substantial performance gap between the reported results for CQL-T (tuned) from (Kumar et al., 2020b) and our own untuned reproduction, CQL-UT (untuned). This gap underscores

Environment	Ours	SAC	CQL-UT	CQL-T
`hopper-medium`	29.11	0.8	25.42	58.0
`halfcheetah-medium`	7.89	-4.3	12.7	44.4

Table 1: Normalized return comparison between our method, our untuned CQL runs (CQL-UT), and SAC + CQL (tuned, CQL-T) baselines as reported in Kumar et al. (2020b).



(a) Comparing q_α

(b) Comparing $Q(s, a)$

Figure 1: SAC vs Conformal SAC

the sensitivity of offline RL algorithms to hyperparameters, particularly in terms of regularization strength and actor update dynamics. We could not execute a full grid search due to computational constraints but implore others to do so in future work (resource permitting).

Conformal Actor-Critic achieves competitive results with both CQL-T and CQL-UT, highlighting the value of uncertainty-aware updtes for improving stability and generalization in offline policy learning.

5.3 Uncertainty Calibration and Interval Dynamics

We track conformal interval width over training both vanilla SAC and Conformal SAC. As shown in Figure 1 (Left) the raw q_α values increase significantly over the course of training, reflecting a growing uncertainty in value estimates as the policy shifts further from the behavior policy. However, our scaled version maintains a much more stable interval, indicating controlled exploration over the course of training. Concurrently, we observe the Q-value estimates under both regimes (see Figure 1 (Right)). In the absence of regularization, Q-values grow rapidly and become unrealistically large (exceeding even expert trajectory rewards) — confirming overestimation bias. In contrast, Conformal Actor-Critic maintains significantly lower Q-values, indicating convservatism and regularization in accordance with calibrated uncertainty. This confirms Theorem 2

5.4 Robustness To OOD Actions

To confirm that Conformal Actor-Critic accurately assigns elevated uncertainty to inputs that diverge from the training distribution, we test the model's behavior on OOD state-action pairs. Specifically, we generate OOD samples using a k-nearest neighbors rejection strategy that selects only those states lying beyond a minimum distance threshold from the offline dataset.

We then compute the confirmal uncertainty ratio – defined as the ratio betwween the confirmal interval width q_α for an OOD sample and the average interval width over the in–distribtion dataset. This provides a scale-invariant way to assess whether the model increases its uncertainty in accordance with distributional shift.

As shown in Figure 3, we observe a consistent positive correlation between uncertainty and distance to the training distribution. On average, OOD samples lie at a distance of 1.622 from the training data (measured in feature space), and are assigned 4 % higher conformal uncertainty on average.

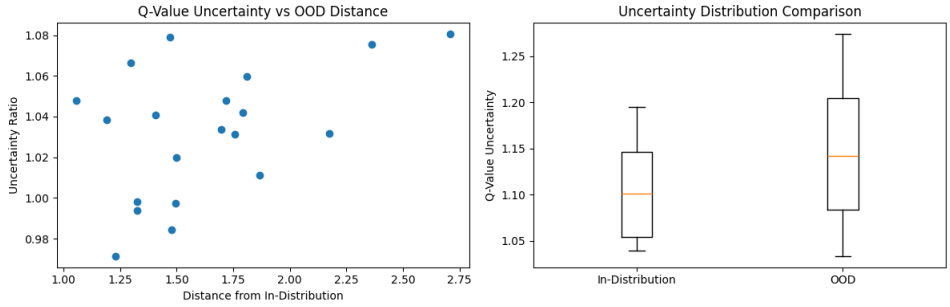


Figure 2: Ratio of conformal interval width $q_\alpha^{\text{OOD}}/q_\alpha^{\text{train}}$ as a function of distance to the training distribution.

Metric	Average Value
Distance from Training Distribution	1.622
Uncertainty Ratio $q_\alpha^{\text{OOD}}/q_\alpha^{\text{train}}$	1.04

Table 2: Aggregate statistics of OOD samples used to evaluate uncertainty robustness.

We measure the model's uncertainty when faced with these OOD states and compare it to its uncertainty with normal states. Our findings show that the model maintains a stable uncertainty level across these tests. On average, the uncertainty ratio stands at approximately 1.04, with the OOD samples averaging a distance of 1.622 from the training distribution. This further supports our claim that conformal uncertainty estimates scale appropriately with distributional shift, even in the absence of explicit OOD labels, providing a useful signal for detecting epistemic uncertainty in offline RL deployments.

5.5 Stability of Actions

Possible Experiment: We evaluated Raw SAC, standard CQL, and our Conformal Actor Critic (CAC) under an offline training regimen of 1 million gradient steps using identical random seeds when possible. During training, we logged evaluation normalized returns at fixed intervals, along with the means and standard deviations of Q-value estimates computed on a fixed batch of state-action pairs. For CAC, we additionally tracked the evolving conformal interval q_α. Post-training, we computed the mean and variance of normalized returns across seeds and analyzed the temporal stability of Q-value estimates.

5.5.1 Robustness to Out-of-Distribution (OOD) Actions

Our model demonstrates robustness against OOD actions through its consistent performance in unfamiliar scenarios. We assess this by using a k-nearest neighbors approach to generate and test OOD samples that maintain a minimum distance from known, in-distribution states. This method helps us verify that the model is fed data outside of its expected distribution.

We measure the model's uncertainty when faced with these OOD states and compare it to its uncertainty with normal states. Our findings show that the model maintains a stable uncertainty level across these tests. On average, the uncertainty ratio stands at approximately 1.04, with the OOD samples averaging a distance of 1.622 from the training distribution.

Metric	Average Value
Average Distance from Training Distribution	1.622
Average Uncertainty Ratio	1.04x

Table 3: Average aggregate statistics of uncertainty ratios and distances for OOD samples

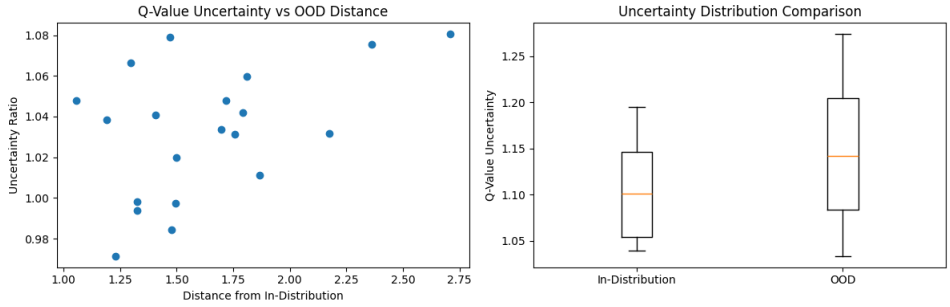


Figure 3: Uncertainty Ratios vs. Distance to Training Distribution

6 Discussion

Our work introduces *Conformal Actor-Critic*, an offline RL framework that integrates conformal prediction into actor-critic and DQN architectures to address extrapolation error. Through theoretical analysis and empirical evaluation, we show how conformal prediction intervals stabilize policy learning, enhance robustness to OOD data, and guide safer decision-making.

We establish finite-sample uncertainty guarantees by constructing theoretical coverage bounds for Q-value estimates. These intervals adapt dynamically based on empirical quantiles, balancing optimism with safety. Our analysis demonstrates that Conformal Actor-Critic is more optimistic yet remains conservatively bounded compared to Conservative Q-Learning (CQL). We operationalize this framework through bang-bang control tasks, where narrowing conformal intervals during training leads to improved policy stability and reduced reactive behavior.

6.1 Limitations and Future Work

While Conformal Actor-Critic shows promising preliminary results, several challenges remain:

- **Hyperparameter Sensitivity:** As observed in our comparison with untuned CQL (CQL-UT), offline RL methods can be highly sensitive to regularization schedules and actor-critic update dynamics. Although conformal intervals help mitigate instability, future work should explore adaptive penalty scaling or automated quantile tuning to further improve robustness.
- **Scalability to High Dimensions:** Our method relies on computing nonconformity scores and empirical quantiles, which may grow expensive in high-dimensional or continuous-action domains. While we use standard SAC infrastructure, extending to more complex domains (e.g., image-based inputs or robotics) may require more efficient calibration methods.
- **Group Structure and Local Calibration:** Our GroupSplitConformal algorithm (which we did not implement due to resource constraints) enables localized uncertainty calibration, but relies on predefined group structures. Learning groupings automatically (e.g., via clustering or representation learning) could yield stronger conditional coverage without domain-specific knowledge.

Conformal Actor-Critic unifies statistical uncertainty quantification with policy optimization in offline RL. The result is a method that is simple, interpretable, and statistically principled—providing formal coverage guarantees without the complexity of ensembles or model-based rollouts. Our results underscore the promise of conformal prediction in reinforcing safety and robustness, particularly for offline decision-making under uncertainty.

We envision future research extending this approach to dynamic and multi-agent environments, integrating adaptive exploration schemes, and scaling to more complex tasks like robotics or autonomous systems—paving the way for safe, uncertainty-aware RL deployments.

ACKNOWLEDGMENTS

We'd like to thank Professor Pratik Chaudhari and Professor Aaron Roth for helpful discussions.

REFERENCES

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *CoRR*, abs/2110.01548, 2021. URL https://arxiv.org/abs/2110.01548.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Gabriel Dulac-Arnold, Nicolas Levine, Daniel Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning, 2021. URL https://arxiv.org/abs/2004.07219.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *CoRR*, abs/2106.06860, 2021. URL https://arxiv.org/abs/2106.06860.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. URL https://arxiv.org/abs/1801.01290.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21810–21823, 2020.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *CoRR*, abs/2110.06169, 2021. URL https://arxiv.org/abs/2110.06169.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020a.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020b.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression, 2017. URL https://arxiv.org/abs/1604.04173.

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

Daniel Liberzon. *Calculus of variations and optimal control theory: a concise introduction*. Princeton university press, 2011.

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. Accelerating online reinforcement learning with offline datasets. *CoRR*, abs/2006.09359, 2020. URL https://arxiv.org/abs/2006.09359.

Marek Petrik, Mohammad Ghavamzadeh, Yinlam Chow, and Morteza Lahijanian. Safe policy improvement by minimizing robust baseline regret. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2281–2289, 2016.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression, 2019. URL https://arxiv.org/abs/1905.03222.

Aaron Roth. Uncertain: Modern topics in uncertainty estimation. Incomplete working draft, PDF notes, 2024.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. doi: 10.1609/aaai.v29i1.9541. URL https://ojs.aaai.org/index.php/AAAI/article/view/9541.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world. *SpringerLink*, 2018. doi: https://doi.org/10.1007-b106715. URL https://link.springer.com/book/10.1007/b106715.

Yue Wu, Shuangfei Zhai, Nitish Srivastava, Joshua M. Susskind, Jian Zhang, Ruslan Salakhutdinov, and Hanlin Goh. Uncertainty weighted actor-critic for offline reinforcement learning. *CoRR*, abs/2105.08140, 2021. URL https://arxiv.org/abs/2105.08140.

Tianhe Yu, Garrett Thomas, Lantao Yu, Aravind Rajeswaran, Chelsea Finn, and Sergey Levine. Mopo: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142, 2020.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, and Paul Christiano. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Appendix

A.1 Proofs

A.1.1 Theorem 1

Similar to Proof 58 in Roth (2024). Sort the data points in D_{cal} in increasing order of their nonconformity scores. Define the index i^* as the index in the dataset with nonconformity score in position $1-\delta$, i.e. $i^* = \lceil (1-\delta)(m+1) \rceil$. Then, since the data points are exchangeable by assumption, the rank of the test nonconformity score is uniformly distributed, and so:

$$Pr(y \in C(s,a)) = \frac{i^*}{m+1} \geq \frac{(1-\delta)(m+1)}{m+1} = 1-\delta$$

The upper bound can be shown by similar analysis:

$$Pr(y \in C(s,a)) = \frac{i^*}{m+1} \leq \frac{(1-\delta)(m+1)+1}{m+1} = 1-\delta+\frac{1}{m+1}$$

A.1.2 Theorem 3

Since $\mathcal{L}(s,a) = f_\theta(s,a) - q_\alpha$ is a lower confidence bound for $Q^\pi(s,a)$, with probability at least $1-\alpha$, it holds that $\mathcal{L}(s,a) \leq Q^\pi(s,a)$ for all (s,a). Therefore, when π_L selects actions based on $\mathcal{L}(s,a)$, it ensures that the expected value $V^{\pi_L}(s)$ does not fall below $V^{\pi_\beta}(s)$ by more than $\epsilon(m,\alpha)$, where $\epsilon(m,\alpha)$ accounts for the finite-sample uncertainty and diminishes as the size of the calibration dataset m increases.

A.1.3 Claim 1: Quantile Consistency

By the DKW inequality, for any $\delta > 0$, setting $t = \sqrt{\frac{\ln(2/\delta)}{2m}}$ gives:

$$\Pr\left(\sup_c |F_\alpha(c) - \hat{F}_\alpha(c)| \leq t\right) \geq 1-\delta.$$

In particular, at $c = \hat{q}_\alpha$ (where $\hat{F}_\alpha(\hat{q}_\alpha) = 1-\alpha$):

$$|F_\alpha(\hat{q}_\alpha) - (1-\alpha)| \leq t.$$

Since $F_\alpha(q_\alpha) = 1-\alpha$, we have:

$$|F_\alpha(\hat{q}_\alpha) - F_\alpha(q_\alpha)| \leq t.$$

Because F_α is monotone, there exists $\epsilon > 0$ such that $|\hat{q}_\alpha - q_\alpha| \leq \epsilon$ implies $|F_\alpha(\hat{q}_\alpha) - F_\alpha(q_\alpha)| \leq t$. Thus, with probability at least $1-\delta$:

$$|\hat{q}_\alpha - q_\alpha| \leq \epsilon.$$

A.1.4 Claim 2: Uniform Convergence

Because each error is bounded by M, define normalized errors:

$$X_i := \frac{e_i}{M}.$$

This ensures that $|X_i| \leq 1$ for all i.

For a fixed state-action pair (s,a), the expected normalized error is $\mathbb{E}[X_i]$. By Hoeffding's inequality, for any $t > 0$:

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mathbb{E}[X_i]\right| \geq t\right) \leq 2e^{-2mt^2}.$$

We want a uniform bound over all $|\mathcal{S}||\mathcal{A}|$ state-action pairs. Applying a union bound, we require that:

$$\Pr\left(\max_{(s,a)}\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mathbb{E}[X_i]\right| < t\right) \geq 1-\delta,$$

Table 4: Hyperparameters for Conformal SAC on D4RL Tasks

Hyperparameter	**Value**
Discount factor γ	0.99
Q-network learning rate	3×10^{-4}
Policy learning rate	3×10^{-5}
Replay buffer capacity	1,000,000
Batch size	256
Training iterations	100,000
Gradient steps per update	1
Soft target update coefficient τ	0.005
Hidden layers	2
Hidden units per layer	256
Activation function	ReLU
Sample frequency	256
Log interval	2000
Conformal calibration ratio	0.1
Conformal update frequency	50 steps
Conformal penalty scale α_q	10
Min log std for policy	-10
Max log std for policy	2

which will hold if:

$$2|\mathcal{S}||\mathcal{A}|e^{-2mt^2} \leq \delta.$$

Solving for t:

$$e^{-2mt^2} \leq \frac{\delta}{2|\mathcal{S}||\mathcal{A}|} \implies -2mt^2 \leq \ln\left(\frac{\delta}{2|\mathcal{S}||\mathcal{A}|}\right).$$

Therefore:

$$t \geq \sqrt{\frac{\ln\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{2m}}.$$

Since $|X_i| \leq 1$, we have $\left|\frac{1}{m}\sum_i X_i - \mathbb{E}[X_i]\right|$ representing the average deviation of the normalized error. Returning to the original scale (i.e. multiplying by M):

$$\max_{(s,a)} |f_\theta(s,a) - Q_\pi(s,a)| = M \max_{(s,a)} \left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mathbb{E}[X_i]\right|.$$

Thus, with probability at least $1-\delta$:

$$\max_{(s,a)} |f_\theta(s,a) - Q_\pi(s,a)| \leq M\sqrt{\frac{2\ln\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{m}}.$$

which gives our desired claim.

A.2 Implementation Details

Our approach integrates an offline conformal prediction framework within the traditional Actor-Critic algorithm:

- **Data Collection:** Data is collected randomly from the environment to ensure a diverse set of training and calibration data.
- **Q-Network Training:** A neural network (termed as Q-Network) predicts Q-values using a dataset of states and actions from the CartPole-v1 environment, based on observed rewards and subsequent states.

- **Conformal Prediction Integration:** During training, a conformal threshold is dynamically calculatedand applied to the updates of Q-values to maintain reliable prediction intervals, stabilizing the policy and ensuring robust decision-making.