MQES: MAX-Q ENTROPY SEARCH FOR EFFICIENT EXPLORATION IN CONTINUOUS REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, the principle of optimism in the face of (aleatoric and epistemic) uncertainty has been utilized to design efficient exploration strategies for Reinforcement Learning (RL). Different from most prior work targeting at discrete action space, we propose a generally information-theoretic exploration principle called Max-Q Entropy Search (MQES) for continuous RL algorithms. MQES formulates the exploration policy to maximize the information about the globally optimal distribution of Q function, which could explore optimistically and avoid over-exploration by recognizing the epistemic and aleatoric uncertainty, respectively. To make MQES practically tractable, we firstly incorporate distributional and ensemble Q function approximations to MQES, which could formulate the epistemic and aleatoric uncertainty accordingly. Then, we introduce a constraint to stabilize the training, and solve the constrained MQES problem to derive the exploration policy in closed form. Empirical evaluations show that MQES outperforms state-of-the-art algorithms on Mujoco environments.

1 INTRODUCTION

In Reinforcement Learning (RL), one of the fundamental problems is exploration-exploitation dilemma, i.e., the agents explore the states with imperfect knowledge to improve future reward or instead maximize the intermediate reward at the perfectly understood states. The main obstacle of designing efficient exploration strategies is how the agents decide whether the unexplored states leading high cumulative reward or not.

Popular exploration strategies, like ϵ-greedy (Sutton & Barto, 1998) and sampling from stochastic policy (Haarnoja et al., 2018), lead to undirected exploration through additional random permutations. Recently, uncertainty of systems are introduced to guide the exploration (Kirschner & Krause, 2018; Mavrin et al., 2019; Clements et al., 2019; Ciosek et al., 2019). Basically, as Moerland et al. (2017) points out, two source of uncertainty exists in the RL system, i.e., epistemic and aleatoric uncertainty. Epistemic uncertainty is also called parametric uncertainty, which is the ambiguity of models arisen from the imperfect knowledge to the environment, and can be reduced with more data. Aleatoric uncertainty is an intrinsic variation associated with the environment, which is caused by the randomness of environment, and is not affected by the model.

By introducing uncertainty, the exploration objectives like Thompson Sampling (TS) (Thompson, 1933; Osband et al., 2016) and Upper Confidence Bound (UCB) (Auer, 2002; Mavrin et al., 2019; Chen et al., 2017) are utilized to guide the exploration in RL. However, since the aleatoric in the RL systems are heteroscedastic, the above methods are not efficient. Hence, Nikolov et al. (2019) proposes novel exploration objective called Information-Directed Sampling (IDS) accounting for epistemic uncertainty and heteroscedastic aleatoric uncertainty. However, these methods (Nikolov et al., 2019; Mavrin et al., 2019; Chen et al., 2017; Osband et al., 2016) can only be applied in the environment with discrete action space.

In this paper, we propose a generally information-theoretic principle called Max-Q Entropy Search (MQES) for off-policy continuous RL algorithms, which is inspired by the entropy-search acquisition functions in the Bayesian Optimization (BO). Further, we combine distributional RL with soft actor-critic method, where the epistemic and aleatoric uncertainty are formulated accordingly.

Then, we incorporate MQES to Distributional Soft Actor-Critic (DSAC) (Ma et al., 2020) method, and show how the MQES utilizes both uncertainty to explore. Finally, our results on Mujoco environments show that our method can substantially outperform alternative state-of-the-art algorithms.

2 Related Work

Efficient exploration can improve the efficiency and performance of RL algorithms. With the increasing emphasis on exploration efficiency, various exploration methods have been developed. One kind of methods use intrinsic motivation to stimulate agent to explore from different perspectives, such as count-based novelty (Martin et al., 2017; Ostrovski et al., 2017; Bellemare et al., 2016; Tang et al., 2017), prediction error (Pathak et al., 2017) and reachability (Savinov et al., 2019). Some recently proposed methods in DRL, originating from tracking uncertainty, do efficient exploration under the principle of OFU (optimism in the face of uncertainty), such as Thompson Sampling (Thompson, 1933; Osband et al., 2016), IDS (Nikolov et al., 2019; Clements et al., 2019) and other customized methods (Moerland et al., 2017; Pathak et al., 2019).

Methods for tracking uncertainty. Bootstrapped DQN (Osband et al., 2016) combines Thompson sampling with value-based algorithms in RL. It is similar to PSRL (Strens, 2000; Osband et al., 2013), and leverages the uncertainty produced by the value estimations for deep exploration. Bootstrapped DQN has become the common baseline for lots of recent works, and also the well-used approach for capturing epistemic uncertainty (Kirschner & Krause, 2018; Ciosek et al., 2019). However, this takes only epistemic uncertainty into account.

Distributional RL approximates the return distribution directly, such as Categorical DQN (C51) (Bellemare et al., 2017), QR-DQN (Dabney et al., 2018b) and IQN (Dabney et al., 2018a). Return distribution can be used to approximate aleatoric uncertainty, but those methods do not take advantage of the return distribution for exploration.

Exploration with two types of uncertainty. Traditional OFU methods either focus only on the epistemic uncertainty, or consider the two kinds of uncertainty as a whole, which can easily lead the naive solution to favor actions with higher variances. To address that, Mavrin et al. (2019) studies how to take advantage of distributions learned by distributional RL methods for efficient exploration under both kinds of uncertainty, proposing Decaying Left Truncated Variance (DLTV).

Nikolov et al. (2019) proposes to use Information Direct Sampling (Kirschner & Krause, 2018) for efficient exploration in RL (IDS for RL), which estimates both kinds of uncertainty via Bootstrapped DQN and return distribution using C51 (Bellemare et al., 2017) respectively. And then it uses IDS to make decision for acting with environment. Clements et al. (2019) addresses computational inefficiencies in IDS for RL, successfully estimating both types of uncertainty only on the expected return, while using two samples from the Bayesian posterior. We refer to the practice of uncertainty estimation in this method as shown in Sec. 4.2. IDS integrates both uncertainty and has made progress on the issue of exploration, but this is a solution on discrete action space. We do focus on how best to exploit both uncertainty for efficient exploration in a continuous action space in our paper.

Optimistic Actor Critic. More closely related to our work is the paper of OAC (Ciosek et al., 2019), which uses epistemic uncertainty to build the upper bound of Q estimation Q^{UB}. OAC is based on Soft Actor-Critic (SAC) (Haarnoja et al., 2018), additionally proposing exploration bonus to facilitate exploration. Despite the advantages that OAC has achieved over SAC, it does not consider the potential impact of the aleatoric uncertainty, which may cause misleading for where the exploration leads.

3 Preliminaries

3.1 Acquisition Functions in Bayesian Optimization

BO methods (Frazier, 2018) are usually used to optimize black-box functions that are costly to evaluate. Typically, these methods derive posteriors over the unknown functions that are updated after each evaluation, and utilize acquisition functions to trade-off exploitation and exploration. Compared with the myopic maximal acquisition functions, e.g., TS and UCB, entropy-search acquisition

functions (Hernandez-Lobato et al., 2015; Wang & Jegelka, 2017), which maximize mutual information between the queried points and the location of the global optimum, could identify good points by using the system uncertainty more efficiently.

3.2 DISTRIBUTIONAL RL

Distributional RL methods study distributions rather than point estimates, which introduces aleatoric uncertainty from distributional perspective. There are different approaches to represent distribution in RL. In our paper, we focus on quantile regression used in QR-DQN (Dabney et al., 2018b), where the randomness of state-action value is represented by the quantile random variable Z. Z maps the state-action pair to a uniform probability distribution supported on z_i, where z_i indicates the value of the corresponding quantile estimates. If τ_i is defined as the quantile fraction, the cumulative probabilities of such quantile distribution is denoted by $F_Z(z_i) = Pr(Z < z_i) = \tau_i = 1/N$ for $i \in 1, ..., N$.

Similar to the Bellman optimality operator in the traditional Q-Learning (Watkins & Dayan, 1992), the distributional Bellman operator $\mathcal{T}_D^\pi$ under policy π is given as:

$$\mathcal{T}_D^\pi Z(s_t, a_t) \overset{D}{=} R(s_t, a_t) + \gamma Z(s_{t+1}, a_{t+1}), \quad a_{t+1} = \pi(s_{t+1}). \tag{1}$$

Notice that this operates on random variables, $\overset{D}{=}$ denotes that distributions on both sides have equal probability laws. Based on the distributional Bellman operator, Dabney et al. (2018b) proposes QR-DQN to train quantile estimations via the quantile regression loss, which is denoted as:

$$\mathcal{L}_{QR}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} [\rho_{\hat{\tau}_i}(\delta_{i,j})] \tag{2}$$

where $\delta_{i,j} = R(s_t, a_t) + \gamma z_j(s_{t+1}, a_{t+1}; \theta) - z_i(s_t, a_t; \theta)$, $\rho_\tau(u) = u * (\tau - \mathbf{1}_{u<0})$, and $\hat{\tau}_i$ means the quantile midpoints, which is defined as $\hat{\tau}_i = \frac{\tau_{i+1}+\tau_i}{2}$.

3.3 DISTRIBUTIONAL SOFT ACTOR-CRITIC METHODS

Following Ma et al. (2020), Distributional RL has been successfully integrated with soft Actor-Critic (SAC) algorithm. Here, considering the maximum entropy RL, the distributional soft Bellman operator $\mathcal{T}_{DS}^\pi$ is defined as follows:

$$\mathcal{T}_{DS}^\pi Z(s, a) \overset{D}{=} R(s_t, a_t) + \gamma[Z(s_{t+1}, \pi(s_{t+1})) - \alpha \log \pi(a_{t+1}|s_{t+1})] \tag{3}$$

where $a' \sim \pi(\cdot|s'), s' \sim \mathcal{P}(\cdot|s, a)$. The quantile regression loss in DSAC is different from original QR-DQN only on the $\delta_{i,j}$ by considering the maximum entropy RL framework. DSAC extends the clipped double Q-Learning proposed on TD3 (Fujimoto et al., 2018) to overcome the overestimation problem. Two Quantile Regression Deep Q Networks have the same structure that are parameterized by $\theta_k, k = 1, 2$. Following the clipped double Q-Learning, the TD-error of DSAC is defined as:

$$y_i^t = \arg\min_{k=1,2} z_i(s_{t+1}, a_{t+1}; \bar{\theta}_k) \tag{4}$$

$$\delta_{i,j}^k = R(s, a) + \gamma[y_i^t - \alpha \arg\log \pi(a_{t+1}|s_{t+1}; \bar{\phi})] - z_j(s_t, a_t; \theta_k) \tag{5}$$

where $\bar{\theta}$ and $\bar{\phi}$ represents their target networks respectively. DSAC has the modified version of critic, while the update of actors is unaffected. It is worth noticing that the state-action value is the minimum value of the expectation on certain distributions, as

$$\begin{aligned} Q(s, a; \theta) &= \arg\min_{k=1,2} Q(s, a; \theta_k) \\ &= \frac{1}{N} \arg\min_{k=1,2} \sum_{i=0}^{N-1} z_i(s, a; \theta_k) \end{aligned} \tag{6}$$

Thus, in DSAC, the original problem aims to maximize the following objective function:

$$\mathcal{J}_\pi(\phi) = \mathbb{E}_{s_t \sim \mathcal{D}, \epsilon \sim \mathcal{N}}[\log \pi(f(s_t, \epsilon_t; \phi)|s_t) - Q(s_t, f(s_t, \epsilon_t; \phi); \theta)], \tag{7}$$

where $\mathcal{D}$ is the replay buffer, $f(s_t, \epsilon_t; \phi)$ means sampling action with re-parameterized policy.

4 ALGORITHM

This paper proposes the MQES, a new exploration principle for continuous RL algorithms, which leverages epistemic and aleatoric uncertainties to explore optimistically and avoid over-exploration. To make MQES practically tractable, distributional and ensemble Q function approximations are introduced to formulate the epistemic and aleatoric uncertainty accordingly. Nevertheless, a constraint is introduced in the MQES to stabilize the training, whereby the exploration policy can be derived in the closed form. All these mechanisms are detailed in the following sections accordingly.

4.1 EXPLORATION STRATEGY: MAX-Q ENTROPY SEARCH

To achieve a better exploration, MQES derives an exploration policy π_E which aims at reducing the epistemic uncertainty and obtain more knowledge of the globally optimal Q function through maximizing the mutual information between (Z^*, π^*) and (Z^{π_E}, π_E). Here, Z^* and Z^{π_E} are random variables measuring the returns (e.g., Q-values) obtained by optimal policy π^* and exploration policy π_E respectively. Specifically, at timestamp t, policy π_E selects action a_t that can maximize the information about the optimal random variable Z^* as follows:

$$\pi_E(a_t|s_t) = \arg\max_{\pi} \mathbf{F}^{\pi}(s_t, a_t), \tag{8}$$

where $\mathbf{F}(\cdot)$ measures the mutual information of the policy π and π^* and can be written as follows:

$$\begin{aligned}\mathbf{F}^{\pi}(s_t, a_t) &= \mathbf{MI}(Z^*(s_t, \pi^*(a_t|s_t)), (Z^{\pi}(s_t, \pi(a_t|s_t)), \pi(a_t|s_t))) \\ &= \mathbf{H}\left[\pi(a_t|s_t)\right] - \mathbf{H}\left[p(a_t|Z^*(s_t, \pi^*(a_t|s_t)))\right].\end{aligned} \tag{9}$$

Here $\mathbf{MI}(\cdot)$ and $\mathbf{H}(\cdot)$ denotes the mutual information and entropy of the random variable, respectively. To obtain behavior policy π_E, we need to measure (1) the posterior probability $p(\cdot)$ in the above equation, and (2) the distributions of $Z^*(\cdot)$ and $Z^{\pi}(\cdot)$. For simplicity, we omit the timestamp t in the following.

To measure the posterior probability $p(a_t|Z^*(s_t, \pi(a_t|s_t)))$, we propose the following proposition.

Proposition 1. *Generally, the posterior probability is as follows :*

$$p(a_t|Z^*(s, \pi(a|s))) \propto \pi(a|s)\Phi_{Z^{\pi}(s,a)}(Z^*(s, a)), \tag{10}$$

where Φ_x is the cumulative distribution function (CDF) of x, Z^* and Z^{π_E} are the random variables, whose distributions describing the randomness of the returns obtained by optimal policy π^* and exploration policy π_E respectively (see proof in Appendix A).

To measure the intractable distribution of Z^* during training, we use the $\hat{Z}^*$ for approximation (i.e., $Z^* \approx \hat{Z}^*$). In general, $\hat{Z}^*$ and Z^{π} are referred to as the optimistic and pessimistic approximator (Mavrin et al., 2019; Chen et al., 2017), respectively, and can be formulated using the uncertainty, which will be detailed in Sec. 4.2.

Therefore, the $\mathbf{F}^{\pi}(s, a)$ in Eq. 9 can be estimated as follows:

$$\mathbf{F}^{\pi}(s, a) \approx \hat{\mathbf{F}}^{\pi}(s, a) = \mathbb{E}_{\pi, \hat{Z}^*}\left[\log \pi(a|s)(G(s, a) - 1) + G(s, a) \log G(s, a)\right], \tag{11}$$

where $G(s, a) = \frac{1}{C} * \Phi_{Z^{\pi}(s,a)}(\hat{Z}^*(s, a))$, and it measures the difference between Z^{π} and $\hat{Z}^*$, i.e., large value of CDF means that $\hat{Z}^*$ is much bigger than the mean of Z^{π}.

4.2 FORMULATION OF Z^{π} AND $\hat{Z}^*$

To derive practically tractable MQES, we formulate the epistemic and aleatoric uncertainty, thereby distributions of $\hat{Z}^*$ and Z^{π} can be estimated. The remaining parts describe how to achieve these two estimations, respectively.

Formulation of $\hat{Z}^*$. In order to formulate the distribution of estimated optimal Q value, a.k.a $\hat{Z}^*$, we firstly estimate its upper confidential bound, denoted by Q^{UB}. Aligned with Clements et al. (2019), we adopt two independent distribution approximators $Z(s, a; \theta_1)$ and $Z(s, a; \theta_2)$ parameterized by θ_1 and θ_2, respectively. Then we measure the epistemic uncertainty first as follows:

$$\sigma_{\text{epistemic}}(s, \pi(a|s); \theta) = \frac{1}{2}\mathbb{E}_{i \sim \mathcal{U}(1,N)}|z_i(s, \pi(a|s); \theta_1) - z_i(s, \pi(a|s); \theta_2)|, \tag{12}$$

where N is the number of quantiles, and $z_i(s,a;\theta)$ is the value of the i-th quantile drawn from $Z(s,a;\theta)$.

Consequently, the upper confidential bound of Q-value is given leveraging the $\sigma_{\text{epistemic}}$ as follows:

$$Q^{\text{UB}}(s,a;\theta) = \mu_Z(s,a;\theta) + \beta\sigma_{\text{epistemic}}(s,a;\theta), \tag{13}$$

where $\mu_Z(s,a;\theta) = \frac{1}{N}\Sigma_{i=1}^{N}\frac{1}{2}\Sigma_{k=1,2}z_i(s,a;\theta_k)$ is the mean estimation over quantile distributions, β determines the magnitude of uncertainty we use. Q^{UB} is commonly considered as a approximation of the optimal Q value in the existing work (Ciosek et al., 2019; Kirschner & Krause, 2018).

Moreover, the aleatoric uncertainty can be derived as follows:

$$\sigma^2_{\text{aleatoric}}(s,\pi(a|s);\theta) = \text{var}_{i\sim\mathcal{U}(1,N)}[\mathbb{E}_{\theta_k}z_i(s,\pi(a|s);\theta_k)]. \tag{14}$$

Inspired by (Wang & Jegelka, 2017), we adopt Q^{UB} and $\sigma^2_{\text{aleatoric}}$ as mean and variation and formulate the Gaussian distribution $\hat{Z}^*$ as follows:

$$\hat{Z}^*(s,a;\theta) \sim \mathcal{N}(Q^{\text{UB}}(s,a;\theta), \sigma^2_{\text{aleatoric}}(s,a;\theta))\mathbf{1}_{\hat{Z}^*\geq Q^{\text{UB}}}, \tag{15}$$

where $\hat{Z}^*$ follows truncated Gaussian distribution ensuring the globally optimal constraint, i.e., $\mathbb{E}[Z^*] = Q^* \geq Q^{UB}$. Nevertheless, since the distributions of Q functions describe the aleatoric uncertainty, we set the variance of $\hat{Z}^*$ as the aleatoric uncertainty obtained from Eq. 14.

Formulation of Z^π. Since the target for critic in the advanced algorithms, like SAC and TD3, is usually estimated pessimistically, we take the pessimistic estimation for Z^π to make MQES compatible with the existing modern RL algorithms. Here we present two modeling approaches: Gaussian and quantile distributions.

Intuitively, we assume Z^π to be the Gaussian distribution with pessimistic estimation as the mean:

$$Z^\pi(s,a;\theta) \sim \mathcal{N}(Q^{\text{LB}}(s,a;\theta), \sigma^2_{\text{aleatoric}}(s,a;\theta)), \tag{16}$$

where $Q^{\text{LB}}(s,a;\theta) = \mu_Z(s,a;\theta) - \beta\sigma_{\text{epistemic}}(s,a;\theta)$, estimates its lower confidential bound.

On the other hand, as the quantile distribution is a value distribution which naturally formulates the underlying aleatoric uncertainty, we can utilize the quantile functions to model the pessimistic quantile distribution directly, breaking the Gaussian assumption above. Specifically, we take the smaller estimates at each quantile, such as

$$Z^\pi(s,a;\theta) = \min_{i=1,2} Z_\tau(s,a;\theta_i). \tag{17}$$

Different from the uni-modal Gaussian distribution, the quantile function is able to represent multi-modal distributions, which is more flexible. As the quantile function represents the inverse function of CDF, meaning that we can easily get a general idea of the properties of this pessimistic quantile distribution.

4.3 MQES-based Exploration For modern RL Algorithms

In this section, we propose a scheme to incorporate exploration policy derived from MQES to existing policy-based algorithms, e.g., SAC and TD3, which renders the stable and well-performed algorithm with more efficient exploration.

First, to obtain exploration policy, we employ a constraint to ensure the difference between the exploration and target policies within a certain range (i.e., $\mathbf{KL}(\pi||\pi_T) \leq \alpha$). The target policy π_T here, we mean the policy learned by any existing policy-based algorithms. It is worth noting that MQES introduces distributional and ensemble critics to the existing framework (e.g., introducing distributional critic to SAC formulates DSAC). Moreover, we utilize the critic of target policy to criticize the exploration policy, i.e., $Z^\pi(s,a) = Z^{\pi_T}(s,a)$ (θ is omitted for simplicity).

Intuitively, introducing the constraint in MQES ensures that the critic of target policy guides the exploration properly and stabilizing the training. Otherwise, the exploration could be ineffective, and the update of target policy can be dramatically bad. Specifically, if the difference between π and π_T are with significant difference, Z^{π_T} could not criticize exploration policy properly, and it

may explore with wrong guidance, where the experiences stored in the replay buffer are with poor quality and the update of target policy fails sequentially.

Second, after introducing the KL constraint, the MQES-based exploration for modern RL algorithms is given as follows:

$$\pi_E(a|s) = \arg\max_{\pi} \quad \hat{\mathbf{F}}^{\pi}(s,a), \\ s.t. \quad \mathbf{KL}(\pi||\pi_T) \leq \alpha, \tag{18}$$

where both the exploration $\pi_E = \mathcal{N}(\mu_E, \Sigma_E)$ and target policy $\pi_T = \mathcal{N}(\mu_T, \Sigma_T)$ are Gaussian distributions. By expanding $\hat{\mathbf{F}}^{\pi}$ linearly, we solve the problem in Eq. 18 using the following proposition:

Proposition 2. *The MQES exploration policy $\pi_E = \mathcal{N}(\mu_E, \Sigma_E)$ derived from Eq. 18 is as follows:*

$$\mu_E = \mu_T + \frac{\sqrt{2\alpha}}{\left\| \mathbb{E}_{\hat{Z}^*}\left[m \odot \frac{\partial \hat{Z}^*(s,a)}{\partial a}|_{a=\mu_T} \right] \right\|_{\Sigma_E}} \Sigma_E \mathbb{E}_{\hat{Z}^*}\left[m \odot \frac{\partial \hat{Z}^*(s,a)}{\partial a}|_{a=\mu_T} \right], \Sigma_E = \Sigma_T. \tag{19}$$

In specific, the i-th element of vector m is $m_i = \log \frac{\Phi(Z^{\pi}(s,\mu_T) \leq \hat{Z}^*(s,\mu_T))}{C\sqrt{(2\pi)\Sigma_{ii}}} + 1$, and n is the action dimension (see proof in Appendix B).

It is worth noting that the expectation against $\mathbb{E}_{\hat{Z}^*}$ can be estimated by sampling, and the unbiased estimation of Eq. 19 is as follows:

$$\mu_E = \mu_T + \frac{\sqrt{2\alpha}}{K} \sum_{i=1}^{K} \frac{1}{\left\| m \odot \frac{\partial \hat{Z}_i^*(s,a)}{\partial a}|_{a=\mu_T} \right\|_{\Sigma_E}} \Sigma_E \left[m \odot \frac{\partial \hat{Z}_i^*(s,a)}{\partial a}|_{a=\mu_T} \right], \tag{20}$$

Algorithm 1 Exploration policy derived from MQES

Initialise: Current state s_t, current value distribution estimators $\theta_k, k = 1, 2$, current policy network ϕ, target policy $\pi_T(a_t|s_t) \sim \mathcal{N}(\mu_T(a_t|s_t;\phi), \sigma_T(a_t|s_t;\phi))$

Output: the MQES exploration policy π_E

1: Calculate $Z(s_t, \mu(a_t|s_t); \theta_k)$, $k = 1, 2$
2: Calculate epistemic uncertainty $\sigma_{\text{epistemic}}(s_t, \mu(a_t|s_t))$ according to Eq. 12
3: Calculate upper bound $Q^{\text{UB}}(s_t, \mu(a_t|s_t))$ using $\sigma_{\text{epistemic}}(s_t, \mu(a_t|s_t))$ according to Eq. 13
4: Calculate aleatoric uncertainty $\sigma^2_{\text{aleatoric}}(s_t, \mu(a_t|s_t))$ according to Eq. 14
5: Construct $\hat{Z}^*(s_t, \mu(a_t|s_t))$ using $Q^{\text{UB}}(s_t, \mu(a_t|s_t))$ and $\sigma^2_{\text{aleatoric}}(s_t, \mu(a_t|s_t))$ (see Eq. 15)
6: Construct $Z^{\pi}(s_t, \mu(a_t|s_t))$ according to Eq. 16 / 17
7: Calculate μ_E using $Z^{\pi}(s_t, \mu(a_t|s_t))$ and $\hat{Z}^*(s_t, \mu(a_t|s_t))$ according to Eq. 20
8: return $\pi_E \sim \mathcal{N}(\mu_E, \sigma_T(a_t|s_t;\phi))$

The above Alg. 1 summarizes the overall procedure of MQES, including the estimation of uncertainty (Line 2,4) and the upper confidential bound Q^{UB} (Line 3), formulation of Z^{π} and $\hat{Z}^*$ (Line 5-6) and exploration policy generation using KL constraint (Line 7). The generated exploration policy can be adopted by any modern policy-based RL algorithms for an more effective exploration.

4.4 Analysis of MQES-based Exploration

This section analytically explains how MQES encourages exploration accounting for the aleatoric and epistemic uncertainty. For simplicity, we assume that the sample number is $K = 1$, and Eq. 20 degrades to:

$$\mu_E = \mu_T + \frac{\sqrt{2\alpha}}{\left\| m \odot \frac{\partial \hat{Z}^*(s,a)}{\partial a}|_{a=\mu_T} \right\|_{\Sigma_E}} \Sigma_E \left[m \odot \frac{\partial \hat{Z}_i^*(s,a)}{\partial a}|_{a=\mu_T} \right], \tag{21}$$

where $m_i = \log \frac{\Phi(Z^{\pi}(s,\mu_T) \leq \hat{Z}^*(s,\mu_T))}{C\sqrt{(2\pi)\Sigma_{ii}}} + 1$.

Take Gaussian MQES for example, we have $Z^\pi(s,a) \sim \mathcal{N}(Q^{LB}(s,a), \sigma^2_{\text{aleatoric}}(s,a))$, and the bias term added to μ_E is decided by the epistemic and aleatoric uncertainty.

Specifically, since epistemic uncertainty is involved in $\hat{Z}^*(s,a)$, the gradient $\frac{\partial \hat{Z}^*(s,a)}{\partial a}$ encourages the optimistic exploration. Epistemic uncertainty-based exploration, can avoid the pessimistic underexploration. The aleatoric uncertainty is introduced by CDF Φ. If we have two state-action pairs, i.e., (s_1, a_1) and (s_2, a_2), and $Z^\pi(s_i, a_i)\mathcal{N}(Q^{LB}(s_i, a_i), \sigma^2_{\text{aleatoric}}(s_i, a_i)), i = 1, 2$, and we assume that $Q^{LB}(s_1, a_1) = Q^{LB}(s_2, a_2)$, $\sigma_1^2 > \sigma_2^2$, and $\hat{Z}^*(s_1, a_1) = \hat{Z}^*(s_2, a_2)$. Obviously, $\Phi(Z^\pi(s_1, a_1) \leq \hat{Z}^*(s_1, a_1)) < \Phi(Z^\pi(s_2, a_2) \leq \hat{Z}^*(s_2, a_2))$, which means that larger aleatoric uncertainty leads to smaller action bias.

Therefore, the MQES encourages the exploration by selecting the action increasing the optimistic value function, and avoid the over-exploration by setting smaller action bias at the state, where the aleatoric uncertainty is high.

5 EXPERIMENTS

MQES is designed for efficient exploration in continuous action space problem in RL, allowing the agent to be aware of explore directions that may lead to higher optimistic value function with smaller aleatoric uncertainty. Comparisons between MQES and state-of-the-art algorithms are conducted to verify the MQES regarding the effectiveness and efficiency. Empirical evaluations show that MQES outperforms state-of-the-art algorithms on a series of continue control tasks.

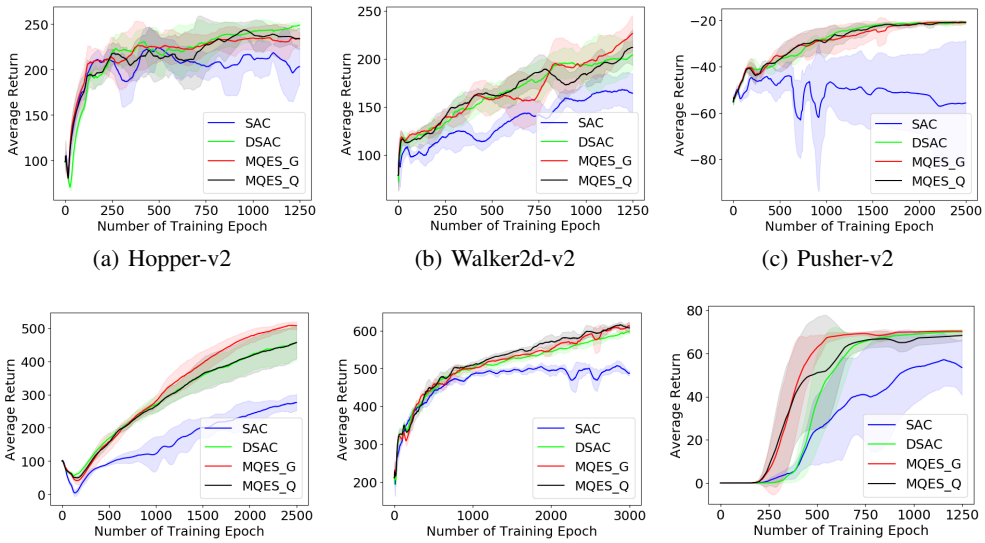
Baselines & Implementation. We compare MQES against SAC (Haarnoja et al., 2018) and its distributional variant DSAC (Ma et al., 2020). Ma et al. (2020) also shows the performance of TD4, which is the distributional extension of TD3 (Fujimoto et al., 2018), and can also be used to capture epistemic and aleatoric uncertainty as is pointed out in Sec. 4.3. However, DSAC outperforms TD4 as shown in Ma et al. (2020), so we evaluate based only on SAC and DSAC, and further implement MQES based on DSAC in order to develop the exploration ability.

The training process of MQES is the same as in DSAC, except for the behavior policy used, while we enrich the experience replay with the data generated by π_E. The pseudo code of the whole process can be found in Appendix C. In order to ensure a fair comparison, the hyper-parameters of DSAC and MQES are the same (see Appendix D). In addition, we have 3 hyper-parameters associated with MQES. The parameter $\sqrt{2\alpha}$ controls the exploration level, and β determines the magnitude of uncertainty we use, and C is the normalization factor.

We implement both approaches for building Z^π as illustrated in Sec. 4.2, and we use MQES_G and MQES_Q to indicate respectively to Gaussian distribution and quantile distribution. We test MQES on 5 tasks in standard Mujoco (Todorov et al., 2012) Environment and we limit the maximum length of each episode to 100. We run 1250 or 3000 epochs for each task, where there are 100 training steps per epoch, with evaluating every epoch, where each evaluation reports the average undiscounted return with no exploration noise.

Evaluations on Mujoco's tasks. The result in Figure 1 (a-e) shows that our method MQES outperforms SAC for all these tasks, and also reaches better performance than DSAC. Our results demonstrate that in complex tasks, such as Humanoid-v2 and Ant-v2, our MQES-based exploration policy performs better, while DSAC suffers from the inefficiency caused by deficient exploration. In Ant-v2, DSAC was overtaken in the early stages of training and then MQES stays ahead. Also in Humanoid-v2, the performance of our algorithm always maintains better than DSAC. In some relatively easier tasks, such as Hopper-v2, it seems that these tasks are not very demanding for exploration, but MQES performs still at a very advanced level. In Walker2d-v2, MQES and DSAC alternated lead until 1000 epochs, after which MQES had a significant improvement. The final results are shown more specifically in the Table 1.

Gaussian and Quantile Z^π. One can find that, expect for Humanoid-v2, there is no absolute superiority between the two modeling approaches. We hypothesis this is because environment of Humanoid-v2 is more complex than others, and more flexible quantile Z^π is needed, which could model the environment more accurately.



(a) Hopper-v2 (b) Walker2d-v2 (c) Pusher-v2

(d) Ant-v2 (e) Humanoid-v2 (f) Sparse Halfcheetah-v2

Figure 1: Training curves on continuous control benchmarks in Mujoco. The x-axis indicates number of training epoch (100 environment steps for each training epoch), while the y-axis is the evaluation result represented by average episode return. The shaded region denotes one standard deviation of average evaluation over 5 seeds. Curves are smoothed uniformly for visual clarity.

Table 1: Average return over 5 seeds with one standard deviation at corresponding training step, i.e., 1.25 x 10^5 million training step for Hopper-v2. The maximum value of each row is shown in bold.

Task	1e5	SAC	DSAC	MQES_G	MQES_Q
Hopper-v2	1.25	203.3 ± 19.9	$\mathbf{248.9 \pm 4.6}$	233.6 ± 7.6	234.2 ± 11.6
Walker-v2	1.25	164.4 ± 20.7	204.2 ± 15.3	$\mathbf{226.6 \pm 18.7}$	212.1 ± 18.7
Pusher-v2	2.5	-55.7 ± 26.9	-21.1 ± 0.6	-21.0 ± 0.6	$\mathbf{-20.8 \pm 0.3}$
Ant-v2	2.5	276.0 ± 24.5	456.5 ± 52.2	$\mathbf{507.5 \pm 12.2}$	456.5 ± 48.4
Humanoid-v2	3.0	487.5 ± 6.0	597.7 ± 4.4	$\mathbf{612.3 \pm 10.5}$	607.3 ± 13.6
Sparse HalfCheetah-v2	1.25	53.5 ± 12.7	70.2 ± 1.1	$\mathbf{70.3 \pm 0.5}$	68.3 ± 2.4

Evaluation on sparse reward environments. To further show the strength of our algorithm regarding exploration, we adjust the standard HalfCheetah-v2 to a sparse version, where the agent no longer receives reward from the environment every step, rather receives reward only when the position reached some defined threshold. As shown in Figure 1(f), SAC performs the worst under this task with sparse environmental rewards, and our MQES-based algorithm converges much faster than DSAC and the final performance is slightly better. It demonstrates that incorporating uncertainty to exploration could render better performance.

6 Conclusion

In this paper, we propose MQES, a generally exploration principle for continuous RL algorithms, which formulates the exploration policy to maximize the information about the globally optimal distribution of Q function. To make MQES practically tractable, we firstly incorporate distributional and ensemble Q function approximations to MQES, which could formulate the epistemic and aleatoric uncertainty accordingly. Secondly, we introduce a constraint to stabilize the training, and solve the constrained MQES problem to derive the exploration policy in closed form. Then, we analyze and show that it explores optimistically and avoid over-exploration by recognizing the epistemic and aleatoric uncertainty, respectively. Empirical evaluations show that MQES works better at the complex environments, where the exploration is needed.

REFERENCES

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, 2002.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pp. 1471–1479, 2016.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 449–458. PMLR, 2017.

Richard Y Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor critic. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 1785–1796, 2019.

William R. Clements, Benoît-Marie Robaglia, Bastien Van Delft, Reda Bahi Slaoui, and Sébastien Toth. Estimating risk and uncertainty in deep reinforcement learning. *CoRR*, abs/1905.09638, 2019.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1104–1113. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 2892–2901. AAAI Press, 2018b.

Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1582–1591. PMLR, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1856–1865. PMLR, 2018.

Jose Miguel Hernandez-Lobato, Michael Gelbart, Matthew Hoffman, Ryan Adams, and Zoubin Ghahramani. Predictive entropy search for bayesian optimization with unknown constraints. volume 37 of *Proceedings of Machine Learning Research*, pp. 1699–1707, 07–09 Jul 2015.

Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pp. 358–384. PMLR, 2018.

Xiaoteng Ma, Qiyuan Zhang, Li Xia, Zhengyuan Zhou, Jun Yang, and Qianchuan Zhao. Distributional soft actor critic for risk sensitive learning. *CoRR*, abs/2004.14547, 2020.

Jarryd Martin, Suraj Narayanan Sasikumar, Tom Everitt, and Marcus Hutter. Count-based exploration in feature space for reinforcement learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2471–2478, 2017.

Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4424–4434. PMLR, 2019.

Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. Efficient exploration with double uncertain value networks. *CoRR*, abs/1711.10789, 2017.

Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-directed exploration for deep reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 3003–3011, 2013.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4026–4034, 2016.

Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2721–2730, 2017.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2778–2787, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5062–5071. PMLR, 2019.

Nikolay Savinov, Anton Raichuk, Damien Vincent, Raphaël Marinier, Marc Pollefeys, Timothy P. Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Malcolm J. A. Strens. A bayesian framework for reinforcement learning. In Pat Langley (ed.), *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pp. 943–950. Morgan Kaufmann, 2000.

R. S. Sutton and A. G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2753–2762, 2017.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109. URL https://doi.org/10.1109/IROS.2012.6386109.

Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient Bayesian optimization. volume 70 of *Proceedings of Machine Learning Research*, pp. 3627–3635, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Christopher J. C. H. Watkins and Peter Dayan. Technical note q-learning. *Mach. Learn.*, 8:279–292, 1992.

A PROOF OF PROPOSITION 1

Proof. The proof is similar to Wang & Jegelka (2017). Since $\hat{Z}^*(s, \pi(a|s))$ is the approximation of the globally optimal Q function, the critic of any policy should be lower than $\hat{Z}^*(s, \pi(a|s))$, i.e.,

$$p(a_t|\hat{Z}^*(s, \pi(a|s))) = \frac{\pi(a|s)\mathbb{E}_{Z^\pi(s,a)}\left[\mathbf{1}_{Z^\pi(s,a)\leq Z^*(s,a)}\right]}{C},$$

where C is the normalization factor, and $\mathbb{E}_{Z^\pi(s,a)}\left[\mathbf{1}_{Z^\pi(s,a)\leq Z^*(s,a)}\right] = \Phi(Z^\pi(s,a) \leq \hat{Z}^*(s, a_t))$ □

B PROOF OF PROPOSITION 2

Proof. Inspired by the methods of multipliers, it is equivalent to maximize objective function $\hat{\mathbf{F}}^\pi(s,a)$ and minimize constraint $\mathbf{KL}(\pi||\pi_T)$ simultaneously. Firstly, we minimize the $\mathbf{KL}(\pi||\pi_T) = \mathbf{KL}(\mathcal{N}(\mu,\Sigma)||\mathcal{N}(\mu_T,\Sigma_T))$ against the variance matrix Σ, i.e.,

$$\min_{\mathbf{\Sigma}} \mathbf{KL}(\mathcal{N}(\mu,\Sigma)||\mathcal{N}(\mu_T,\Sigma_T)). \tag{22}$$

The optimal solution for problem (22) is

$$\Sigma_E = \Sigma_T. \tag{23}$$

Then, $\hat{\mathbf{F}}^\pi(s,a)$ is expanded linearly around $a_t = \mu_T$:

$$\hat{\mathbf{F}}^\pi(s,a) \approx a^T \nabla_a \hat{\mathbf{F}}^\pi(s,a)|_{a=\mu_T} + const \tag{24}$$

$$= a^T \left[\hat{m} \odot \frac{\partial \hat{Z}^*(s,a)}{\partial a}|_{a=\mu_T}\right] + const, \tag{25}$$

where the element of vector $\hat{m} = \{\hat{m}_i\}_{i=1}^n$ is

$$\hat{m}_i = \frac{1}{C}\phi(Z^\pi(s,\mu_T) \leq \hat{Z}^*(s,\mu_T))(\log \frac{\Phi(Z^\pi(s,\mu_T) \leq \hat{Z}^*(s,\mu_T))}{C\sqrt{(2\pi)\Sigma_{ii}}} + 1), \tag{26}$$

n is the dimension of action, and $\phi(x)$ is the probability distribution function (pdf). Then, problem (18) is reformulated as:

$$\begin{aligned} \max_{\mu} \quad & \mathbb{E}_{\hat{Z}^*}\left\{\mu^T\left[m \odot \frac{\partial \hat{Z}^*(s,a)}{\partial a}|_{a=\mu_T}\right]\right\} \\ s.t. \quad & \frac{1}{2}(\mu-\mu_T)^T\Sigma^{-1}(\mu-\mu_T) \leq \alpha, \end{aligned} \tag{27}$$

Then, the Lagrange function of problem (27) is given as:

$$L(\mu) = \mathbb{E}_{\hat{Z}^*}\left\{\mu^T\left[\hat{m} \odot \frac{\partial \hat{Z}^*(s,a)}{\partial a}|_{a=\mu_T}\right]\right\} - \lambda\left[\frac{1}{2}(\mu-\mu_T)^T\Sigma^{-1}(\mu-\mu_T) - \alpha\right]. \tag{28}$$

According to the KKT condition, we derive the following equations: First, from the stationary, i.e., $\nabla_\mu L(\mu) = m \odot \frac{\partial \hat{Z}^*(s,a)}{\partial a}|_{a=\mu_T} - \lambda\Sigma^{-1}(\mu-\mu_T) = 0$, we get

$$\mu = \mu_T + \frac{1}{\lambda}\Sigma\mathbb{E}_{\hat{Z}^*}\left[\hat{m} \odot \frac{\partial \hat{Z}^*(s,a)}{\partial a}|_{a=\mu_T}\right], \tag{29}$$

where $\lambda > 0$. Secondly, since $\lambda > 0$, the constraint is active, i.e., $(\mu-\mu_T)^T\Sigma^{-1}(\mu-\mu_T) = 2\alpha$. Together with Eq. 29, we get the following equation:

$$\lambda = \frac{\left\|\mathbb{E}_{\hat{Z}^*}\left[\hat{m} \odot \frac{\partial \hat{Z}^*(s,a)}{\partial a}|_{a=\mu_T}\right]\right\|_\Sigma}{\sqrt{2\alpha}}. \tag{30}$$

Finally, (19) is obtained by plugging (30) to (29). □

C ALGORITHM 2: MQES FOR DSAC

D HYPER-PARAMETERS SETTING

Algorithm 2 MQES for DSAC

Initialise: Value networks θ_1, θ_2, policy network ϕ and their target networks $\bar{\theta}_1$, $\bar{\theta}_2$, $\bar{\phi}$, quantiles number N, target smoothing coefficient (τ), discount (γ), an empty replay pool $\mathcal{D}$

for each iteration **do**
  for each environmental step **do**
    $a_t \sim \pi_E(a_t, s_t)$ according to Algorithm 1
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1})\}$
  end for
  for each training step **do**
    for i = 1 to N **do**
      for j = 1 to N **do**
        calculate $\delta^k_{i,j}, k = 1, 2$, following Eq. 5
      end for
    end for
    Calculate $\mathcal{L}_{QR}(\theta_k), k = 1, 2$ using $\delta^k_{i,j}$ following Eq. 2
    Update θ_k with $\nabla \mathcal{L}_{QR}(\theta_k)$
    Calculate $\mathcal{J}_\pi(\phi)$, following Eq. 7
    Update ϕ with $\nabla \mathcal{J}_\pi(\phi)$
  end for
  Update target value network with $\bar{\theta}_k \leftarrow \tau\theta_k + (1-\tau)\bar{\theta}_k, k = 1, 2$
  Update target policy network with $\bar{\phi} \leftarrow \tau\phi + (1-\tau)\bar{\phi}$
end for

Table 2: MQES parameters

	Parameter		Value
Training	discount		0.99
	Target smoothing coefficient	τ	5e-3
	Learning rate		3e-4
	Optimizer		Adam
	Replay buffer size		10e6
	Batch size		256
	Quantiles number		20
	Action range		[-1, 1]
	Environment steps per epoch		100
Exploration	Exploration ratio	$\sqrt{2\alpha}$	0.8
	Uncertainty ratio	β	1.6
	Normalization factor	C	0.5