Table 6: Hyperparameter settings

Type	Hyperparameter	TD3-SMR	DARC-SMR	SAC-SMR	REDQ-SMR	MAMC
Shared	#Actors (N_A)	1	2	1	1	10
	#Critics (N_C)	2	2	2	10	10
	Discount factor	0.99	0.99	0.99	0.99	0.99
	Actor learning rate	3.0E-4	3.0E-4	3.0E-4	3.0E-4	1.0E-4[1]
	Critic learning rate	3.0E-4	3.0E-4	3.0E-4	3.0E-4	3.0E-4
	Optimizer	Adam	Adam	Adam	Adam	Adam
	Batch size ($N_{\mathcal{B}}$)	256	256	256	256	256
	Actor target	v	v	-	-	-
	Critic target	v	v	v	v	v
	Soft update ratio (τ)	5.0E-3	5.0E-3	5.0E-3	5.0E-3	5.0E-3
	SMR ratio (M)	10	10	10	10	10
	Warm-up steps	5k	5k	5k	5k	5k
	Delayed update (d)	2	1	1	10	1
Deterministic	Exploration noise	$\mathcal{N}(0, 0.1)$	$\mathcal{N}(0, 0.1)$	-	-	$\mathcal{N}(0, 0.1)$
	Target policy noise	$\mathcal{N}(0, 0.2)$	$\mathcal{N}(0, 0.2)$	-	-	$\mathcal{N}(0, 0.1)$
	Noise clip	$[-0.5, 0.5]$	$[-0.5, 0.5]$	-	-	-
Stochastic	Temperature (α)	-	-	Tuned[2]	Adaptive	-
	Log std. clip	-	-	$[-20, 2]$	$[-20, 2]$	-
Specific	Weighting coef. (ν)	-	Tuned[3]	-	-	-
	Regularization (λ)	-	5.0E-3	-	-	-
	Target entropy	-	-	-	Tuned[4]	-
	Ensemble subset size	-	-	-	2	-
	Quantile (q)	-	-	-	-	0.2

A Experimental Settings in Detail

This section gives detailed experimental settings adopted in this study. The code along with the instructions containing the exact command and environment needed to run to reproduce the results, and the followed licenses are available at `https://github.com/NobodyAcademic/MAMC`.

A.1 Hyperparameter Settings

Table 6 compiles the hyperparameter settings for the three deterministic-policy-based (TD3-SMR, DARC-SMR, and MAMC) and two stochastic-policy-based (SAC-SMR and REDQ-SMR) methods. Most of the settings follow the original suggestions in the non-SMR version. In the shared hyperparameters, the number of actors and critics in the MAMC are both set to 10, which equals to the number of critics in REDQ-SMR. In addition, the DARC-SMR, SAC-SMR, and MAMC have no delayed update for each actor, whilst TD3-SMR and REDQ-SMR has a delayed update of 2 and 10, respectively. Furthermore, SAC-SMR, REDQ-SMR, and the MAMC do not consider the utilization of actor target when calculating the TD target. Noteworthily, this study sets a low actor learning rate for the MAMC since it has no delayed update and actor target. All the test methods have an SMR ratio of 10. As REDQ-SMR has considered SMR technique, its UTD ratio is set to 1 for a fair comparison.

For hyperparameters considered in deterministic-policy-based methods, the proposed MAMC adds noise to actors when exploration and calculation of target values with the same distribution, while TD3-SMR and DARC-SMR considered larger noise when computing the target values than exploration. Also, the MAMC has no noise clip for simplicity. As for hyperparameters leveraged in stochastic-policy-based methods, the SAC-SMR set a small

[1] Without delayed update and target actor, the MAMC adopts a small learning rate.

[2] SAC set the α to 0.05 for Humanoid, and 0.2 for the others.

[3] DARC set the ν to 0.15 for Hopper, 0.25 for Ant, and 0.1 for the others.

[4] REDQ set target entropy to -1 for Hopper, -2 for Humanoid, -3 for HalfCheetah and Walker, and -4 for Ant.

temperature for Humanoid, and a large one for the others, and the REDQ-SMR considered an adaptive control of temperature.

Some hyperparameters are exploited in a specific method. DARC-SMR fine-tuned weighting coefficient ν for different environment, and considered a regularization coefficient for similarity of two critics. REDQ-SMR also fine-tuned the target entropy for each environment, and set the ensemble subset size to 2. For the MAMC, the number of actors and critics are both set to 10, and the quantile parameter q is set to 0.2.

A.2 System Configuration

All the experiments are conducted on a server with Intel Xeon W7-2475X CPU (with 2.6 GHz clock rate, 20 cores and 40 hyperthreads), two NVIDIA RTX 4090 GPU cards (each with 24GB memory), and 128 GB main memory.

A.3 MuJoCo

The properties of the selected environments in MuJoCo [31] are listed as follows:

- Hopper-v5
  - Appearance: 2D single-leg hopping robot
  - Simulation: kangaroo hopping
  - State: 11-dimensional random vector $s \in \mathbb{R}^{11}$, includes position and velocity information of various body parts
  - Action: 3-dimensional random vector $a \in [-1, 1]^3$, corresponding to torque control of three hinge joints
- HalfCheetah-v5
  - Appearance: 2D bipedal robot
  - Simulation: cheetah running
  - State: 17-dimensional random vector $s \in \mathbb{R}^{17}$, includes joint angles, angular velocities, and body linear velocity
  - Action: 6-dimensional random vector $a \in [-1, 1]^6$, corresponding to torque control of six hinge joints
- Walker2d-v5
  - Appearance: 2D bipedal walking robot
  - Simulation: human walking
  - State: 17-dimensional random vector $s \in \mathbb{R}^{17}$, includes position and velocity information of various body parts
  - Action: 6-dimensional random vector $a \in [-1, 1]^6$, corresponding to torque control of six hinge joints
- Ant-v5
  - Appearance: 3D quadrupedal robot
  - Simulation: ant walking
  - State: 105-dimensional random vector $s \in \mathbb{R}^{105}$, includes position, velocity, and angle information of various body parts
  - Action: 8-dimensional random vector $a \in [-1, 1]^8$, corresponding to torque control of eight hinge joints
- Humanoid-v5
  - Appearance: 3D bipedal humanoid robot
  - Simulation: complex human-like locomotion and balancing
  - State: 348-dimensional random vector $s \in \mathbb{R}^{348}$, includes joint angles, velocities, torso orientation, and center of mass information
  - Action: 17-dimensional random vector $a \in [-0.4, 0.4]^{17}$, corresponding to torque control of 17 motor joints

B Proof of Theorems

Theorem 1. *The variance of target values obtained by multiple actors are less than that using a single actor*

$$\mathbb{V}[\hat{V}_A(s'; C')] \leq \mathbb{V}[\hat{V}_\phi(s'; C')]\,. \tag{16}$$

Proof. Assume that the distribution of $\{\hat{V}_{\phi_i}(s';C')\}_{1\leq i\leq N_A}$ are not skewed (symmetric), we have:

$$\begin{aligned}
\mathbb{V}_{s'\sim S}[\hat{V}_A(s';C')] &= \mathbb{V}[\mathrm{Med}(\{\hat{V}_{\phi_i}(s';C')\}_{1\leq i\leq N_A})] \\
&= \mathbb{V}[\mathbb{E}_{\phi_i\in A}[\hat{V}_{\phi_i}(s';C')]] \\
&= \mathbb{V}[N_A^{-1}\textstyle\sum_{\phi_i\in A}\hat{V}_{\phi_i}(s';C')] \\
&= N_A^{-2}\textstyle\sum_{\phi_i\in A}\mathbb{V}[\hat{V}_{\phi_i}(s';C')] \\
&\leq N_A^{-1}\mathbb{V}[\hat{V}_{\phi_{\max}}(s';C')] \\
&\leq \mathbb{V}[\hat{V}_{\phi_{\min}}(s';C')] \\
&\leq \mathbb{V}[\hat{V}_{\phi}(s';C')] \\
&\leq \mathbb{V}[\hat{V}_{\phi_{\max}}(s';C')]\,. \qquad (17)
\end{aligned}$$

The inequality is always satisfied comparing to $\phi = \phi_{\max}$. For generalization to any arbitrary $\phi \geq \phi_{\min}$, the ratio of maximum to minimum variance are within some bound

$$\mathbb{V}_{\phi_{\max}}/\mathbb{V}_{\phi_{\min}} \leq \epsilon_A\,, \qquad (18)$$

where $\epsilon_A = N_A$ serves as a constraint. Also, it is apparent that the larger the N_A the easier the satisfaction of the constraint on the ratio.

□

Theorem 2. *The variance of target values obtained by multiple critics are less than using a single critic*

$$\mathbb{V}[\hat{V}_\phi(s';C')] \leq \mathbb{V}[\hat{V}_\phi(s';\theta')]\,. \qquad (19)$$

Proof. Assume that the q-th quantile among critic targets C' is c_q times their expectation:

$$\begin{aligned}
\hat{V}_\phi(s';C') &= \mathrm{Quantile}_q(\{Q_{\theta'_j}(s',\pi_\phi(s'))\}_{1\leq j\leq N_C}) \\
&= c_q\mathbb{E}_{\theta'\in C'}[Q_{\theta'}(s',\pi_\phi(s'))]\ \exists\ c_q\in\mathbb{R}
\end{aligned}\,, \qquad (20)$$

and thus the following equation proves the theorem:

$$\begin{aligned}
\mathbb{V}_{s'\sim S}[\hat{V}_\phi(s';C')] &= \mathbb{V}[c_q\mathbb{E}_{\theta'\in C'}[Q_{\theta'}(s',\pi_\phi(s'))]] \\
&= c_q^2\mathbb{V}[N_C^{-1}\textstyle\sum_{\theta'\in C'}Q_{\theta'}(s',\pi_\phi(s'))] \\
&= c_q^2N_C^{-2}\textstyle\sum_{\theta'\in C'}\mathbb{V}[Q_{\theta'}(s',\pi_\phi(s'))] \\
&\leq c_q^2N_C^{-1}\mathbb{V}[Q_{\theta'_{\max}}(s',\pi_\phi(s'))] \\
&\leq \mathbb{V}[Q_{\theta'_{\min}}(s',\pi_\phi(s'))] \\
&= \mathbb{V}[\hat{V}_\phi(s';\theta'_{\min})] \\
&\leq \mathbb{V}[\hat{V}_\phi(s';\theta')] \\
&\leq \mathbb{V}[\hat{V}_\phi(s';\theta'_{\max})]\,. \qquad (21)
\end{aligned}$$

This theorem holds when the ratio of maximum to minimum variance are within some bound

$$\mathbb{V}_{\theta'_{\max}}/\mathbb{V}_{\theta'_{\min}} \leq \epsilon_C\,, \qquad (22)$$

subject to

$$\epsilon_C = c_q^{-2}N_C\,. \qquad (23)$$

The bound ϵ_C can be viewed as a constraint of SAMC to be more stable than SASC. From the above equation, it is obvious that the intensity of the constraint is proportional to the coefficient c_q and is inverse proportional to the number of critics.

□

For proving the next theorems, this study first introduces two lemmas.

Lemma 1. The target values among multiple actors are in between the minimum and maximum of target values for a single actor

$$\mathbb{E}[\hat{V}_{\phi_{\min}}(s';C)] \leq \mathbb{E}[\hat{V}_A(s';C)] \leq \mathbb{E}[\hat{V}_{\phi_{\max}}(s';C)]\,. \tag{24}$$

Proof. The lemma holds owing to the following inequality:

$$\hat{V}_{\phi_{\min}}(s';C) \leq \hat{V}_A(s';C) \leq \hat{V}_{\phi_{\max}}(s';C)\,. \tag{25}$$

□

Lemma 2. The target values among multiple critics are in between the minimum and maximum of target values for a single critic

$$\mathbb{E}[\hat{V}_A(s';\theta_{\min})] \leq \mathbb{E}[\hat{V}_A(s';C)] \leq \mathbb{E}[\hat{V}_A(s';\theta_{\max})]\,. \tag{26}$$

Proof. Similarly, the inequality holds with

$$\hat{V}_A(s';\theta_{\min}) \leq \hat{V}_A(s';C) \leq \hat{V}_A(s';\theta_{\max})\,. \tag{27}$$

□

Theorem 3. *The estimation error of MAMC is between the estimation error of multiple actors with minimum and maximum critics*

$$\mathcal{E}_{A,Q_{\theta_{\min}}} \leq \mathcal{E}_{A,C} \leq \mathcal{E}_{A,Q_{\theta_{\max}}}\,. \tag{28}$$

Proof. The proof is similar to the one given in [21]:

$$\begin{aligned}
\mathcal{E}_{A,Q_{\theta_{\min}}} &= \mathbb{E}[\hat{V}_A(s';\theta_{\min})] - \mathbb{E}[V_{\phi^*}(s')] \\
&\leq \mathbb{E}[\hat{V}_A(s';C)] - \mathbb{E}[V_{\phi^*}(s')] \\
&= \mathcal{E}_{A,C} \\
&\leq \mathbb{E}[\hat{V}_A(s';\theta_{\max})] - \mathbb{E}[V_{\phi^*}(s')] \\
&= \mathcal{E}_{A,Q_{\theta_{\max}}}\,.
\end{aligned} \tag{29}$$

□

Theorem 4. *The estimation error of MAMC is between the estimation error of multiple critics with minimum and maximum actors*

$$\mathcal{E}_{\pi_{\phi_{\min}},C} \leq \mathcal{E}_{A,C} \leq \mathcal{E}_{\pi_{\phi_{\max}},C}\,. \tag{30}$$

Proof. Similar derivation can be applied:

$$\begin{aligned}
\mathcal{E}_{\pi_{\phi_{\min}},C} &= \mathbb{E}[\hat{V}_{\phi_{\min}}(s';C)] - \mathbb{E}[V_{\phi^*}(s')] \\
&\leq \mathbb{E}[\hat{V}_A(s';C)] - \mathbb{E}[V_{\phi^*}(s')] \\
&= \mathcal{E}_{A,C} \\
&\leq \mathbb{E}[\hat{V}_{\phi_{\max}}(s';C)] - \mathbb{E}[V_{\phi^*}(s')] \\
&= \mathcal{E}_{\pi_{\phi_{\max}},C}\,,
\end{aligned} \tag{31}$$

and the theorem is proved.

□

C Additional Experimental Results

Additional experimental results and further analysis are given in the following subsections.

Table 7: Wilcoxon signed rank test for TD3 and DARC compared with the MAMC at early (100k), middle (200k), and late stage (300k). The win/tie/lose denotes the number of environments that the MAMC is significantly superior (+), equal (~), and inferior (-) to a corresponding test method.

Stage	p-value	TD3-SMR	DARC-SMR	SAC-SMR	REDQ-SMR
100k	Hopper-v5	5.27E-02 (~)	2.44E-02 (−)	1.61E-01 (~)	2.44E-02 (-)
	HalfCheetah-v5	1.38E-01 (~)	6.88E-01 (~)	6.15E-01 (~)	4.61E-01 (~)
	Walker2d-v5	4.88E-03 (+)	3.22E-02 (+)	1.37E-02 (+)	3.85E-01 (~)
	Ant-v5	9.77E-04 (+)	9.77E-04 (+)	2.93E-03 (+)	4.23E-01 (~)
	Humanoid-v5	9.77E-03 (+)	9.67E-02 (~)	2.44E-02 (+)	5.00E-01 (~)
Summary (win/tie/lose)		3/2/0	2/2/1	3/2/0	0/4/1
200k	Hopper-v5	6.54E-02 (~)	9.67E-02 (~)	1.86E-02 (-)	9.77E-04 (-)
	HalfCheetah-v5	5.77E-01 (~)	1.88E-01 (~)	5.27E-02 (~)	5.27E-02 (~)
	Walker2d-v5	3.48E-01 (~)	2.78E-01 (~)	9.67E-02 (~)	2.46E-01 (~)
	Ant-v5	9.77E-04 (+)	9.77E-04 (+)	1.95E-03 (+)	9.67E-02 (~)
	Humanoid-v5	4.88E-03 (+)	1.88E-01 (~)	4.20E-02 (+)	5.00E-01 (~)
Summary (win/tie/lose)		2/3/0	1/4/0	2/2/1	0/4/1
300k	Hopper-v5	1.38E-01 (~)	5.39E-01 (~)	2.78E-01 (~)	8.01E-02 (~)
	HalfCheetah-v5	6.88E-01 (~)	1.38E-01 (~)	5.27E-02 (~)	9.77E-03 (-)
	Walker2d-v5	3.13E-01 (~)	4.20E-02 (-)	1.61E-01 (~)	2.46E-01 (~)
	Ant-v5	9.77E-04 (+)	1.95E-03 (+)	1.95E-03 (+)	8.01E-02 (~)
	Humanoid-v5	1.37E-02 (+)	5.77E-01 (~)	4.61E-01 (~)	4.20E-02 (-)
Summary (win/tie/lose)		2/3/0	1/3/1	1/4/0	0/3/2

C.1 Statistical Analysis

Table 7 compiles the Wilcoxon signed rank test for TD3 and DARC compared with the MAMC at early (100k), middle (200k), and late stage (300k). The win/tie/lose denotes the number of environments that the MAMC is significantly superior (+), equal (~), and inferior (-) to a corresponding test method. The MAMC betters TD3-SMR, DARC-SMR, and SAC-SMR at all three stage. In addition, the MAMC is comparable to REDQ-SMR at early and middle stages, yet is inferior to the REDQ-SMR at late stage.

C.2 Quantile Value Comparison

Figure 4 draws the average return against environment steps for MAMC with different quantile parameters $q \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ on HalfCheetah-v5 and Walker2d-v5. On HalfCheetah-v5, the MAMCs with $q = 0.1$, 0.3, and 0.5 are better, while on Walker2d-v5 the MAMCs with $q = 0.2$, and 0.3 performs nicer. The quantile parameter highly hinges

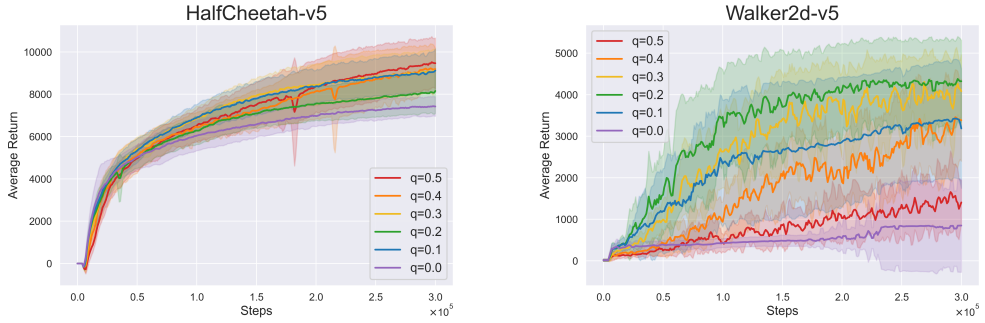

Figure 4: Average return against environmental steps for MAMC with different quantile parameters $q \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ on HalfCheetah-v5 and Walker2d-v5

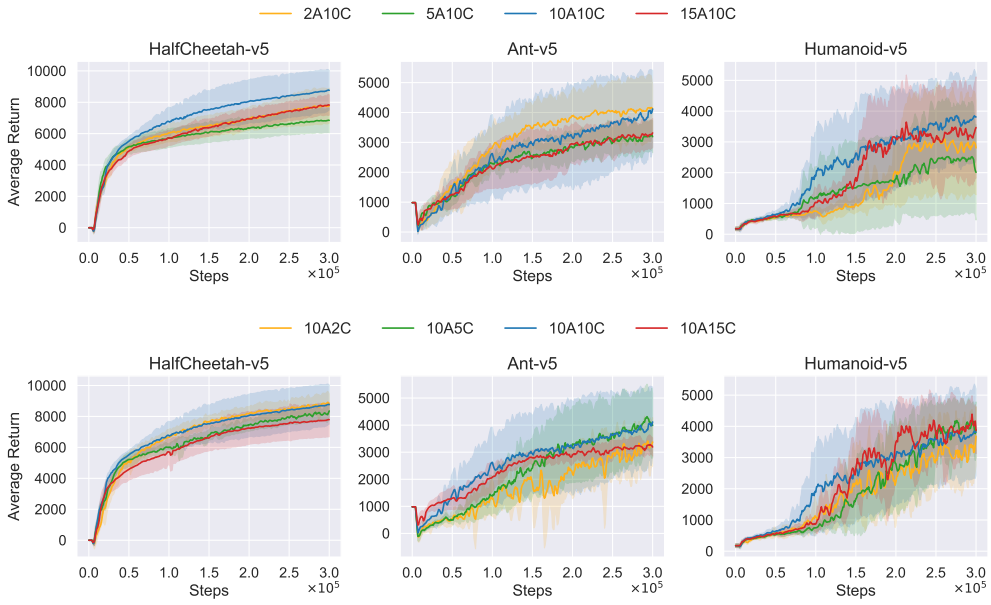

Figure 5: Average return against environmental steps for MAMC with different number of actors $N_A \in \{2, 5, 10, 15\}$ and critics $N_C \in \{2, 5, 10, 15\}$ on HalfCheetah-v5, Ant-v5, and Humanoid-v5 over five trials

on the environmental preference of optimism or pessimism. From the experimental results, this study would suggest setting $q \in [0.2, 0.3]$ for better robustness.

C.3 The number of Actors and Critics

Figure 5 depicts the average return against environmental steps for MAMC with different number of actors $N_A \in \{2, 5, 10, 15\}$ and critics $N_C \in \{2, 5, 10, 15\}$ on HalfCheetah-v5, Ant-v5, and Humanoid-v5 over five trials. For setting the number of actors, the MAMC with $N_A = 10$ performs best, and the performance deteriorates as the number of actors grows to 15 or shrinks to 5 and 2. By varying the number of critics, the MAMC with $N_C = 10$ provides the most robust results on the three environments, in comparison to the other three values. Hence, this study suggests taking 10 actors and critics for the MAMC.