Learning to Walk a Tripod Mobile Robot Using Nonlinear Soft Vibration Actuators with Entropy Adaptive Reinforcement Learning: Supplementary Material

Jae In Kim*, Mineui Hong*, Kyungjae Lee, DongWook Kim, Yong-Lae Park, and Songhwai Oh

In this supplementary material, we provide proofs of lemmas and theorems in the main paper and comparisons of our algorithm with other actor-critic algorithms in MuJoCo simulator. This material consists of three sections. In Section I, we first derive the trust region temperature adaptation. We also provide the proof of the optimality of adaptive soft actor-critic algorithm in Section II. Finally, the experimental results on four MuJoCo simulation tasks are shown in Section III.

I. Trust Region Temperature Adaptation

In this section, we present the derivation of the proposed trust region temperature adaptation, which finds a new temperature α_{m+1} by solving an optimization problem below.

$$\begin{aligned} \underset{\alpha_{m+1}}{\text{maximize}} \quad & \mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}},a\sim\pi_{\alpha_m}}\left[\frac{\pi_{\alpha_{m+1}}(a_t|s_t)}{\pi_{\alpha_m}(a_t|s_t)}\hat{Q}^{\pi_{\alpha_m}}(s_t,a_t)\right], \\ \text{subject to} \quad & \mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}}}\left[D_{KL}(\pi_{\alpha_m}(\cdot|s_t)||\pi_{\alpha_{m+1}}(\cdot|s_t))\right] \leq \delta. \end{aligned} \tag{1}$$

First, we prove that the quadratic approximation of the KL-divergence term in Equation (1) is computed as,

$$D_{KL}(\pi_{\alpha_m}(\cdot|s)||\pi_{\alpha_{m+1}}(\cdot|s)) \approx \frac{(\alpha_{m+1}-\alpha_m)^2}{2{\alpha_m}^4}\mathbb{E}_{a\sim\pi_{\alpha_m}}\left[\left(\hat{Q}^{\pi_{\alpha_m}}(s,a)-\hat{V}^{\pi_{\alpha_m}}(s)\right)^2\right]. \tag{2}$$

Proof: First, note that Q^π_α can be decomposed as,

$$Q^\pi_\alpha(s,a) = Q^\pi(s,a) + \alpha\mathbb{E}_{s'\sim P}\left[\gamma\mathcal{H}^\infty_\pi(s')\right], \tag{3}$$

where,

$$\begin{aligned} Q^\pi_\alpha(s,a) &:= \mathbb{E}_{\tau\sim P,\pi}\left[r(s_0,a_0)+\sum_{t=1}^{\infty}\gamma^t\left(r(s_t,a_t)+\alpha\mathcal{H}\left(\pi\left(\cdot|s_t\right)\right)\right)|s_0=s,a_0=a\right] \\ Q^\pi(s,a) &:= \mathbb{E}_{\tau\sim P,\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)|s_0=s,a_0=a\right] \\ \mathcal{H}^\infty_\pi(s) &:= \mathbb{E}_{\tau\sim P,\pi}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{H}(\pi(\cdot|s_t))|s_0=s\right]. \end{aligned} \tag{4}$$

Then, since $Q^\pi(s,a)$ and $\mathcal{H}^\infty_\pi(s)$ are independent of α for fixed π,

$$\frac{d}{d\alpha}\left(\frac{1}{\alpha}Q^\pi_\alpha(s,a)\right) = -\frac{1}{\alpha^2}Q^\pi(s,a). \tag{5}$$

Now, let π_{α_m} denotes a given old policy, and assume that it has been trained for enough iterations with a temperature α_m, and has converged to $\pi^*_{\alpha_m}$. Then the following soft Bellman optimality equation holds.

$$\pi_{\alpha_m}(a|s) = \exp\left(\frac{1}{\alpha_m}\left(\hat{Q}^{\pi_{\alpha_m}}_{\alpha_m}(s,a)-\alpha_m\log\int_{\mathcal{A}}\exp(\frac{1}{\alpha_m}\hat{Q}^{\pi_{\alpha_m}}_{\alpha_m}(s,a'))da'\right)\right). \tag{6}$$

Where the soft policy iteration, $\mathcal{I}_\alpha$, is defined as,

$$\mathcal{I}_\alpha\pi := \exp\left(\frac{1}{\alpha}\left(\hat{Q}^\pi_\alpha(s,a)-\alpha\log\int_{\mathcal{A}}\exp(\frac{1}{\alpha}\hat{Q}^\pi_\alpha(s,a'))da'\right)\right), \tag{7}$$

we can define the new policy $\pi_{\alpha_{m+1}}$ as, $\pi_{\alpha_{m+1}} = \mathcal{I}_{\alpha_{m+1}}\pi_{\alpha_m}$:

$$\pi_{\alpha_{m+1}}(a|s) = \exp\left(\frac{1}{\alpha_{m+1}}\left(\hat{Q}^{\pi_{\alpha_m}}_{\alpha_{m+1}}(s,a)-\alpha_{m+1}\log\int_{\mathcal{A}}\exp(\frac{1}{\alpha_{m+1}}\hat{Q}^{\pi_{\alpha_m}}_{\alpha_{m+1}}(s,a'))da'\right)\right), \tag{8}$$

and,

$$\frac{d\pi_{\alpha_{m+1}}(a|s)}{d\alpha_{m+1}} = \frac{\pi_{\alpha_{m+1}}(a|s)}{{\alpha_{m+1}}^2}\left(-\hat{Q}^{\pi_{\alpha_m}}(s,a) + \frac{\int_{\mathcal{A}} \hat{Q}^{\pi_{\alpha_m}}(s,a')\exp(\frac{1}{\alpha_{m+1}}\hat{Q}^{\pi_{\alpha_m}}_{\alpha_{m+1}}(s,a'))da'}{\int_{\mathcal{A}}\exp(\frac{1}{\alpha_{m+1}}\hat{Q}^{\pi_{\alpha_m}}_{\alpha_{m+1}}(s,a'))da'}\right)$$
$$= \frac{\pi_{\alpha_{m+1}}(a|s)}{{\alpha_{m+1}}^2}\left(-\hat{Q}^{\pi_{\alpha_m}}(s,a) + \int_{\mathcal{A}} \hat{Q}^{\pi_{\alpha_m}}(s,a')\pi_{\alpha_{m+1}}(a'|s)da'\right). \tag{9}$$

Now, using Taylor expansion, we can approximate $D_{KL}(\pi_{\alpha_m}(\cdot|s)||\pi_{\alpha_{m+1}}(\cdot|s))$ as,

$$\begin{aligned} D_{KL}(\pi_{\alpha_m}(\cdot|s)||\pi_{\alpha_{m+1}}(\cdot|s)) = &\left[D_{KL}(\pi_{\alpha_m}(\cdot|s)||\pi_{\alpha_{m+1}}(\cdot|s))\right]_{\alpha_{m+1}=\alpha_m} \\ &+ (\alpha_{m+1}-\alpha_m)\left[\frac{d}{d\alpha_{m+1}}D_{KL}(\pi_{\alpha_m}(\cdot|s)||\pi_{\alpha_{m+1}}(\cdot|s))\right]_{\alpha_{m+1}=\alpha_m} \\ &+ \frac{(\alpha_{m+1}-\alpha_m)^2}{2}\left[\frac{d^2}{{d\alpha_{m+1}}^2}D_{KL}(\pi_{\alpha_m}(\cdot|s)||\pi_{\alpha_{m+1}}(\cdot|s))\right]_{\alpha_{m+1}=\alpha_m}, \end{aligned} \tag{10}$$

for $|\alpha_{m+1}-\alpha_m| \ll 1$.

It is straightforward that $[\pi_{\alpha_{m+1}}(a|s)]_{\alpha_{m+1}=\alpha_m} = \pi_{\alpha_m}(a|s)$, therefore,

$$\left[D_{KL}(\pi_{\alpha_m}(\cdot|s)||\pi_{\alpha_{m+1}}(\cdot|s))\right]_{\alpha_{m+1}=\alpha_m} = 0. \tag{11}$$

Also, we can show that $\left[\frac{d}{d\alpha_{m+1}}D_{KL}(\pi_{\alpha_m}(\cdot|s)||\pi_{\alpha_{m+1}}(\cdot|s))\right]_{\alpha_{m+1}=\alpha_m} = 0$.

$$\begin{aligned} &\left[\frac{d}{d\alpha_{m+1}}D_{KL}(\pi_{\alpha_m}(\cdot|s)||\pi_{\alpha_{m+1}}(\cdot|s))\right]_{\alpha_{m+1}=\alpha_m} \\ &= \left[\frac{d}{d\alpha_{m+1}}\int_{\mathcal{A}}\pi_{\alpha_m}(a|s)\log\frac{\pi_{\alpha_m}(a|s)}{\pi_{\alpha_{m+1}}(a|s)}da\right]_{\alpha_{m+1}=\alpha_m} \\ &= \left[\int_{\mathcal{A}} -\frac{\pi_{\alpha_m}(a|s)}{\pi_{\alpha_{m+1}}(a|s)}\frac{d\pi_{\alpha_{m+1}}(a|s)}{d\alpha_{m+1}}da\right]_{\alpha_{m+1}=\alpha_m} \\ &= \left[\int_{\mathcal{A}}\frac{\pi_{\alpha_m}(a|s)}{{\alpha_{m+1}}^2}\left(\hat{Q}^{\pi_{\alpha_m}}(s,a) - \int_{\mathcal{A}}\hat{Q}^{\pi_{\alpha_m}}(s,a')\pi_{\alpha_{m+1}}(a'|s)da'\right)da\right]_{\alpha_{m+1}=\alpha_m} \\ &= \int_{\mathcal{A}}\frac{\pi_{\alpha_m}(a|s)}{{\alpha_{m+1}}^2}\left(\hat{Q}^{\pi_{\alpha_m}}(s,a) - \int_{\mathcal{A}}\hat{Q}^{\pi_{\alpha_m}}(s,a')\pi_{\alpha_m}(a'|s)da'\right)da = 0. \end{aligned} \tag{12}$$

We now compute $\left[\frac{d^2}{{d\alpha_{m+1}}^2}D_{KL}(\pi_{\alpha_{m+1}}(\cdot|s)||\pi_{\alpha_m}(\cdot|s))\right]_{\alpha_{m+1}=\alpha_m}$ as,

$$\begin{aligned} &\left[\frac{d^2}{{d\alpha_{m+1}}^2}D_{KL}(\pi_{\alpha_{m+1}}(\cdot|s)||\pi_{\alpha_m}(\cdot|s))\right]_{\alpha_{m+1}=\alpha_m} \\ &= \left[\frac{d^2}{{d\alpha_{m+1}}^2}\int_{\mathcal{A}}\pi_{\alpha_m}(a|s)\log\frac{\pi_{\alpha_m}(a|s)}{\pi_{\alpha_{m+1}}(a|s)}da\right]_{\alpha_{m+1}=\alpha_m} \\ &= \left[\frac{d}{d\alpha_{m+1}}\int_{\mathcal{A}}\frac{\pi_{\alpha_m}(a|s)}{{\alpha_{m+1}}^2}\left(\hat{Q}^{\pi_{\alpha_m}}(s,a) - \int_{\mathcal{A}}\hat{Q}^{\pi_{\alpha_m}}(s,a')\pi_{\alpha_{m+1}}(a'|s)da'\right)da\right]_{\alpha_{m+1}=\alpha_m} \\ &= \left[\frac{1}{{\alpha_{m+1}}^2}\int_{\mathcal{A}}\pi_{\alpha_m}(a|s)\left(\int_{\mathcal{A}} -\hat{Q}^{\pi_{\alpha_m}}(s,a')\frac{d\pi_{\alpha_{m+1}}(a'|s)}{d\alpha_{m+1}}da'\right)da\right]_{\alpha_{m+1}=\alpha_m} \\ &= \left[\frac{1}{{\alpha_{m+1}}^2}\int_{\mathcal{A}} -\hat{Q}^{\pi_{\alpha_m}}(s,a')\frac{d\pi_{\alpha_{m+1}}(a'|s)}{d\alpha_{m+1}}da'\right]_{\alpha_{m+1}=\alpha_m} \\ &= \left[\frac{1}{{\alpha_{m+1}}^2}\int_{\mathcal{A}}\hat{Q}^{\pi_{\alpha_m}}(s,a')\left(\hat{Q}^{\pi_{\alpha_m}}(s,a') - \int_{\mathcal{A}}\hat{Q}^{\pi_{\alpha_m}}(s,a'')\pi_{\alpha_{m+1}}(a''|s)da''\right)\pi_{\alpha_{m+1}}(a'|s)da'\right]_{\alpha_{m+1}=\alpha_m} \\ &= \frac{1}{{\alpha_m}^4}\int_{\mathcal{A}}\hat{Q}^{\pi_{\alpha_m}}(s,a)\left(\hat{Q}^{\pi_{\alpha_m}}(s,a) - \int_{\mathcal{A}}\hat{Q}^{\pi_{\alpha_m}}(s,a')\pi_{\alpha_m}(a'|s)da'\right)\pi_{\alpha_m}(a|s)da \\ &= \frac{1}{{\alpha_m}^4}\int_{\mathcal{A}}\left(\hat{Q}^{\pi_{\alpha_m}}(s,a) - \int_{\mathcal{A}}\hat{Q}^{\pi_{\alpha_m}}(s,a')\pi_{\alpha_m}(a'|s)da'\right)^2\pi_{\alpha_m}(a|s)da. \end{aligned} \tag{13}$$

Finally the quadratic approximation of $D_{KL}(\pi_{\alpha_{m+1}}(\cdot|s)||\pi_{\alpha_m}(\cdot|s))$ can be computed as,

$$D_{KL}(\pi_{\alpha_{m+1}}(\cdot|s)||\pi_{\alpha_m}(\cdot|s)) \approx \frac{(\alpha_{m+1}-\alpha_m)^2}{2{\alpha_m}^4}\mathbb{E}_{a\sim\pi_{\alpha_m}}\left[\left(\hat{Q}^{\pi_{\alpha_m}}(s,a)-\hat{V}^{\pi_{\alpha_m}}(s)\right)^2\right], \tag{14}$$

where, $\hat{V}^{\pi}(s)=\int_{\mathcal{A}}\hat{Q}^{\pi}(s,a)\pi(a|s)da$.

Now, note that

$$\begin{aligned}
&\left[\frac{d}{d\alpha_{m+1}}\mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}}a\sim\pi_{\alpha_m}}\left[\frac{\pi_{\alpha_{m+1}}(a_t|s_t)}{\pi_{\alpha_m}(a_t|s_t)}\hat{Q}^{\pi_{\alpha_m}}(s_t,a_t)\right]\right]_{\alpha_{m+1}=\alpha_m}\\
&=\mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}}a\sim\pi_{\alpha_m}}\left[\frac{1}{{\alpha_m}^2}\left(-\hat{Q}^{\pi_{\alpha_m}}(s_t,a_t)+\int_{\mathcal{A}}\hat{Q}^{\pi_{\alpha_m}}(s_t,a)\pi_{\alpha_{m+1}}(a|s_t)da\right)\hat{Q}^{\pi_{\alpha_m}}(s_t,a_t)\right]\\
&=-\mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}}a\sim\pi_{\alpha_m}}\left[\frac{1}{{\alpha_m}^2}\left(\hat{Q}^{\pi_{\alpha_m}}(s_t,a_t)-\hat{V}^{\pi_{\alpha_m}}(s_t)\right)^2\right]<0,
\end{aligned} \tag{15}$$

which means that, $\mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}}a\sim\pi_{\alpha_m}}\left[\frac{\pi_{\alpha_{m+1}}(a_t|s_t)}{\pi_{\alpha_m}(a_t|s_t)}\hat{Q}^{\pi_{\alpha_m}}(s_t,a_t)\right]$ increases as α_{m+1} decreases. Therefore the solution of Equation (1) appears at the equality of KL constraints, i.e., $\mathbb{E}_{\rho_{\pi_{\alpha_m}}}\left[D_{KL}(\pi_{\alpha_m}(\cdot|s_t)||\pi_{\alpha_{m+1}}(\cdot|s_t))\right]=\delta$ and $\alpha_{m+1}<\alpha_m$. Then, we can compute the new temperature α_{m+1} as,

$$\begin{aligned}
&\alpha_{m+1}=\alpha_m-{\alpha_m}^2\sqrt{\frac{2\delta}{\mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}}a\sim\pi_{\alpha_m}}\left[\hat{A}^{\pi_{\alpha_m}}(s_t,a_t)^2\right]}},\\
&\text{where, } \hat{A}^{\pi_{\alpha_m}}(s,a)=\hat{Q}^{\pi_{\alpha_m}}(s,a)-\hat{V}^{\pi_{\alpha_m}}(s).
\end{aligned} \tag{16}$$

II. Optimality of Adaptive Soft Actor-Critic

Algorithm 1 Adaptive Soft Actor-Critic

Initialize parameter vectors $\psi,\bar{\psi},\theta_i,\phi,\lambda,\omega$, entropy coefficient α, and replay buffer D.
for each iteration **do**
  for each environment steps **do**
    Sample a transition $\{s_t,a_t,r(s_t,a_t),s_{t+1}\}$, and store it in the replay buffer D.
  end for
  for each gradient steps **do**
    Minimize $J_{V^\alpha}(\psi)$, $J_{Q^\alpha}(\theta_{1,2})$, $J_Q(\mu)$, $J_V(\omega)$, and $J_\pi(\phi)$ using stochastic gradient descent.
    $\bar{\psi}\leftarrow(1-\tau)\bar{\psi}+\tau\psi$
  end for
  if π_ϕ converges **then**
    Update α with trust region method
  end if
end for

In this section, we prove that adaptive soft actor-critic (ASAC) using the trust region method (Equation 16), can find π^*, an optimal policy of the orignal MDP.

First, we define the performance of a policy π, as the expected discounted reward sum:

$$J(\pi)=\mathbb{E}_{\tau\sim P,\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]. \tag{17}$$

Also, the optimal policy of the original MDP, π^* can be defined as $\pi^*=\arg\max_\pi J(\pi)$, and let $\mathcal{H}_\pi^\infty(s)$ denotes the expected discounted entropy sum of a policy π, from an initial state s:

$$\mathcal{H}_\pi^\infty(s)=\mathbb{E}_{\tau\sim P,\pi}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{H}(\pi(\cdot|s_t))|s_0=s\right]. \tag{18}$$

Then, we can show that if $\{\alpha_m\}$ converges to zero, $\pi^*_{\alpha_m}$, the optimal policy of the soft MDP with a temperature α_m, converges to π^*.

Lemma 1. *Consider a decreasing sequence of entropy temperature* $\{\alpha_m\}$*, such that* $\alpha_m > 0$ *and* $\lim_{m\to\infty}\alpha_m = 0$*, and corresponding* $\pi^*_{\alpha_m}$*. Then,* $\lim_{m\to\infty}\pi^*_{\alpha_m} = \pi^*$.

Proof: Since $\pi^* = \arg\max_\pi J(\pi)$, it is straightforward that $J(\pi^*_{\alpha_m}) \le J(\pi^*)$. Also, since π^* is a deterministic policy, $\mathcal{H}(\pi^*(\cdot|s)) = 0$ for all $s \in \mathcal{S}$. Then, by the definition of $\pi^*_{\alpha_m}$,

$$\begin{aligned} J(\pi^*) = J(\pi^*) + \alpha_m \mathbb{E}_{s\sim d}\left[\mathcal{H}^\infty_{\pi^*}(s)\right] &\le J(\pi^*_{\alpha_m}) + \alpha_m \mathbb{E}_{s\sim d}\left[\mathcal{H}^\infty_{\pi^*_{\alpha_m}}(s)\right] \\ &\le J(\pi^*_{\alpha_m}) + \alpha_m \mathbb{E}_{s\sim d}\left[\sum_{t=0}^{\infty}\gamma^t\left(-\log\frac{1}{|\mathcal{A}|}\right)\right] \\ &= J(\pi^*_{\alpha_m}) + \frac{\alpha_m}{1-\gamma}\log|\mathcal{A}|. \end{aligned} \tag{19}$$

where $|\mathcal{A}|$ is a cardinality of the action space $\mathcal{A}$. Therefore, we can know that $J(\pi^*_{\alpha_m})$ is bounded as,

$$J(\pi^*) - \frac{\alpha_m}{1-\gamma}\log|\mathcal{A}| \le J(\pi^*_{\alpha_m}) \le J(\pi^*). \tag{20}$$

Since $\alpha_m > 0$ and $\lim_{m\to\infty}\alpha_m = 0$, for all $\epsilon > 0$, there exists $M \in \mathbb{N}$, such that, $m > M \Rightarrow 0 < \alpha_m < \frac{(1-\gamma)\epsilon}{\log|\mathcal{A}|}$. Then, $m > M \Rightarrow J(\pi^*) - \epsilon \le J(\pi^*_{\alpha_m}) \le J(\pi^*)$. Therefore, $\lim_{m\to\infty} J(\pi^*_{\alpha_m}) = J(\pi^*)$, and by the definition of π^*, $\lim_{m\to\infty}\pi^*_{\alpha_m} = \pi^*$.

Now, we show that a sequence of temperatures $\{\alpha_m\}$, which is made by the trust region method, converges to zero. First, we assume that $\mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}}a\sim\pi_{\alpha_m}}\left[\hat{A}^{\pi_{\alpha_m}}(s_t,a_t)^2\right]$ is bounded as,

$$0 < L < \mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}}a\sim\pi_{\alpha_m}}\left[\hat{A}^{\pi_{\alpha_m}}(s_t,a_t)^2\right] < U. \tag{21}$$

Then, we can show the following lemma.

Lemma 2. *Let* $\{\alpha_m\}$ *be a sequence of entropy temperatures, made by Equation 16 from an initial temperature* α_0*, such that,* $0 < \alpha_0 < \sqrt{\frac{L}{2\delta}}$*. Then,* $\alpha_m > \alpha_{m+1} > 0$ *for all* m*, and* $\lim_{m\to\infty}\alpha_m = 0$.

Proof: Since, $\alpha_m{}^2\sqrt{\frac{2\delta}{\mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}}a\sim\pi_{\alpha_m}}\left[\hat{A}^{\pi_{\alpha_m}}(s_t,a_t)^2\right]}}$ is always greater then zero, it is straightforward that $\{\alpha_m\}$ is a decreasing sequence. Therefore, $\alpha_m < \alpha_0 < \sqrt{\frac{L}{2\delta}}$ for all m, and then, if we assume α_m is greater than zero, we can show that α_{m+1} is also greater than zero.

$$\begin{aligned} \alpha_{m+1} &= \alpha_m - \alpha_m{}^2\sqrt{\frac{2\delta}{\mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}}a\sim\pi_{\alpha_m}}\left[\hat{A}^{\pi_{\alpha_m}}(s_t,a_t)^2\right]}} \\ &> \alpha_m - \alpha_m{}^2\sqrt{\frac{2\delta}{L}} \\ &> \alpha_m - \alpha_m = 0. \end{aligned} \tag{22}$$

Therefore, $\alpha_m > 0$ for all m, by mathematical induction and the only remaining part is to show $\lim_{m\to\infty}\alpha_m = 0$.

As shown above, $\{\alpha_m\}$ is a decreasing sequence and has a lower bound zero, then there exists α, which is the infimum of α_m, and α_m converges to α as $m \to \infty$.

$$\exists\alpha, \text{ such that, } \alpha = \inf\{\alpha_m\} \ge 0, \quad \lim_{m\to\infty}\alpha_m = \alpha. \tag{23}$$

Now assume $\alpha > 0$. Then, for $\epsilon = \frac{\alpha^2}{2}\sqrt{\frac{2\delta}{U}}$, there exist $M \in \mathbb{N}$, such that $m \ge M \Rightarrow \alpha < \alpha_m < \alpha + \epsilon$. Then,

$$\begin{aligned} \alpha_{M+1} &= \alpha_M - \alpha_M{}^2\sqrt{\frac{2\delta}{\mathbb{E}_{s\sim\rho_{\pi_{\alpha_m}}a\sim\pi_{\alpha_m}}\left[\hat{A}^{\pi_{\alpha_m}}(s_t,a_t)^2\right]}} \\ &< \alpha_M - \alpha_M{}^2\sqrt{\frac{2\delta}{U}} \\ &< (\alpha+\epsilon) - \alpha^2\sqrt{\frac{2\delta}{U}} = \alpha - \frac{\alpha^2}{2}\sqrt{\frac{2\delta}{U}} < \alpha. \end{aligned} \tag{24}$$

It is contradiction to the definition of $\alpha = \inf\{\alpha_m\}$. Thus, $\alpha = 0$, and therefore $\lim_{m\to\infty} \alpha_m = 0$.

Then now, we can finally show the optimality of adaptive soft actor-critic, using Lemma 1 and 2.

Theorem 1. *Consider a sequence of temperatures $\{\alpha_m\}$ made by Equation 16. Then, repeated application of adaptive soft policy iteration with $\{\alpha_m\}$, from any initial policy π_0, converges to an optimal policy π^*.*

Proof: From Lemma 2, the given sequence of temperature $\}\alpha_m\}$ is a decreasing sequence which converges to zero. Since adaptive soft actor-critic with a sequence of the temperature $\{\alpha_m\}$ updates the policy to sequentially converge to the $\pi^*_{\alpha_m}$, the policy can finally converges to π^* by Lemma 1.

III. Simulation Experiments of ASAC

We also verify that the proposed learning algorithm can be used for general RL problems, by evaluating the algorithm on four MuJoCo simulation tasks (HalfCheetah-v2, Pusher-v2, Ant-v2, and Humanoid-v2), and comparing it with other actor-critic algorithms.

A. Implementation Details for Adaptive Soft Actor-Critic

The hyperparameters for implementation of ASAC for simulation experiments are detailed in the table below.

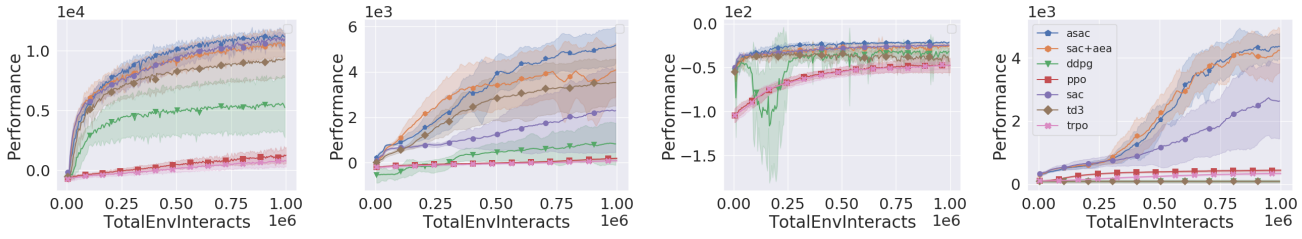
Parameter	Value
Threshold δ for Trust Reigon Method	0.01
Optimizer	Adam in TensorFlow
Learning rate	$5e^{-4}$
Discount factor	0.99
Replay buffer size	$1e^6$
Number of Minimum samples in buffer	$1e^5$
Number of Hidden Layers	2
Number of Hidden units	$[300, 400]$
Activation function	ReLu
Number of samples in minibatch	100
Moving average ratio	0.005
Seeds	0 10 20 30 40 50 60 70 80 90

Environment	Degree of Freedom	Initial Entropy temperature, α_0
HalfCheetah-v2	6	0.2
Pusher-v2	7	0.2
Ant-v2	8	0.2
Humanoid-v2	17	0.05

Also, to determine whether to update the temperature or not, we compare the changes of $J_\pi(\phi)$ for every 1000 update steps, and ASPI decide to reduce the temperature if $\frac{J_\pi(\phi_{old}) - J_\pi(\phi_{new})}{J_\pi(\phi_{old})} < 0.0001$.

B. Experimental Results

In this section, we present simulation experimental results on four MuJoCo tasks. Figure 1 compares ASAC with all the baseline methods (SAC, SAC-AEA, TD3, DDPG, PPO, and TRPO). ASAC shows the highest expected return and the smallest variance in all the tasks.



(a) HalfCheetah-v2 (b) Ant-v2 (c) Pusher-v2 (d) Humanoid-v2

Fig. 1. Comparison to all the baseline methods on four MuJoCo tasks. All the figures share the legend.

Figure 2 and 3 show the expected return of SAC and ASAC on four Mujoco tasks with different entropy temperatures (or different initial values of entropy temperature). As shown in Figure 3, ASAC adapts to all the different tasks and initial