Supplementary Material for Generalized Tsallis Entropy Reinforcement Learning and Its Application to Soft Mobile Robots

Kyungjae Lee[1], Sungyub Kim[2], Sungbin Lim[3], Sungjoon Choi[1], Mineui Hong[1], Jaein Kim[4], Yong-Lae Park[4] and Songhwai Oh[1]
[1]Dept. of Electrical and Computer Engineering, ASRI, Seoul National University,
[2]Graduate School of AI, Korea Advanced Institute of Science and Technology (KAIST),
[3]Dept. of Industrial Engineering, Ulsan National Institute of Science and Technology (UNIST),
[4]Dept. of Mechanical and Aerospace Engineering, Seoul National University
Email: {kyungjae.lee,mineui.hong,sungjoon.choi}@rllab.snu.ac.kr,sungyub.kim@kaist.ac.kr,
sungbin@unist.ac.kr,{snu08mae,ylpark,songhwai}@snu.ac.kr

APPENDIX

We show that the Tsallis entropy is a concave function over the distribution P and has the maximum at an uniform distribution. Note that this is an well known fact, but, we restate it to make the manuscript self-contained.

Proposition 1. *Assume that $\mathcal{X}$ is a finite space. Let P is a probability distribution over $\mathcal{X}$. If $q > 0$, then, $S_q(P)$ is concave with respect to P.*

Proof: Let us consider the function $f(x) = -x \ln_q(x)$ defined over $(x > 0)$. Second derivative of $d^2 f(x)/dx^2$ is computed as

$$\frac{d^2 f(x)}{dx^2} = -q x^{q-2} < 0 \;\; (x > 0, q > 0).$$

Thus, $f(x)$ is a concave function. Now, using this fact, we show that the following inequality holds. For $\lambda_1, \lambda_2 \geq 0$ such that $\lambda_1 + \lambda_2 = 1$, and probabilities P_1 and P_2,

$$\begin{aligned} S_q(\lambda_1 P_1 + \lambda_2 P_2) &= \sum_x -(\lambda_1 P_1(x) + \lambda_2 P_2(x)) \ln_q(\lambda_1 P_1(x) + \lambda_2 P_2(x)) \\ &< \sum_x -\lambda_1 P_1(x) \ln_q(P_1(x)) - \lambda_2 P_2(x) \ln_q(P_2(x)) \\ &= \lambda_1 S_q(P_1) + \lambda_2 S_q(P_2). \end{aligned}$$

Consequently, $S_q(P)$ is concave with respect to P. ■

Proposition 2. *Assume that $\mathcal{X}$ is finite space. Then, $S_q(P)$ is maximized when P is a uniform distribution, i.e., $P = 1/|\mathcal{X}|$ where $|\mathcal{X}|$ is the number of elements in $\mathcal{X}$.*

Proof: We would like to employ the KKT condition on the following optimization problem.

$$\max_{P \in \Delta} S_q(P) \tag{1}$$

where $\Delta = \{P | P(x) \geq 0, \sum_x P(x) = 1\}$ is a probability simplex. Since $\mathcal{X}$ is finite, the optimization variables are probability mass defined over each element. The KKT condition of 1 is

$$\begin{aligned} &\forall x \in \mathcal{X}, \left. \frac{\partial \left(S_q(\pi) - \sum_x \lambda^\star(x) P(x) - \mu^\star \left(1 - \sum_x P(x)\right)\right)}{\partial P(x)} \right|_{P(x) = P^\star(x)} \\ &\qquad = -\ln_q(P^\star(x)) - (P^\star(x))^{q-1} - \lambda^\star(x) + \mu^\star \\ &\qquad = -q \ln_q(P^\star(x)) - 1 - \lambda^\star(x) + \mu^\star = 0 \\ &\forall x \in \mathcal{X}, 0 = 1 - \sum_x P^\star(x), P^\star(x) \geq 0 \\ &\forall x \in \mathcal{X}, \lambda^\star(x) \leq 0 \\ &\forall x \in \mathcal{X}, \lambda^\star(x) P^\star(x) = 0 \end{aligned}$$

where $\lambda^\star$ and $\mu^\star$ are the Lagrangian multipliers for constraints in Δ. First, let us consider $P^\star(x) > 0$. Then, $\lambda^\star(x) = 0$ from the last condition (complementary slackness). The first condition implies

$$P^\star(x) = \exp_q\left(\frac{\mu^\star - 1}{q}\right).$$

Hence, $P^\star(x)$ has constant probability mass which means $P^\star(x) = 1/|S|$ where $S = \{x|P^\star(x) > 0\}$. The optimal value is $S_q(P^\star) = -\ln_q(1/|S|)$. Since $-\ln_q(x)$ is a monotonically decreasing function, $|S|$ should be the largest number as possible as it can be. Hence, $S = \mathcal{X}$ and $P^\star(x) = 1/|\mathcal{X}|$. ■

Now, we prove the property of q-maximum which is defined by

$$q\text{-}\max_x(f(x)) \triangleq \max_{P\in\Delta}\left[\mathop{\mathbb{E}}_{X\sim P}[f(X)] + S_q(P)\right] \tag{2}$$

Theorem 1. *For any function $f(x)$ defined on finite input space $\mathcal{X}$, the q-maximum satisfies the following inequalities.*

$$q\text{-}\max_x\left(f(x)\right) + \ln_q\left(1/|\mathcal{X}|\right) \le \max_x(f(x)) \le q\text{-}\max_x\left(f(x)\right) \tag{3}$$

where $|\mathcal{X}|$ is a cardinality.

Proof: First, consider the lower bound. Let Δ be a probability simplex. Then,

$$\begin{aligned} q\text{-}\max_x(f(x)) &= \max_{P\in\Delta}\left[\mathop{\mathbb{E}}_{X\sim P}[f(X)] + S_q(P)\right] \le \max_{P\in\Delta}\mathop{\mathbb{E}}_{X\sim P}[f(X)] + \max_{P\in\Delta} S_q(P) \\ &= \max_x(f(x)) - \ln_q\left(\frac{1}{|\mathcal{X}|}\right) \end{aligned} \tag{4}$$

where $S_q(P)$ has the maximum at an uniform distribution.

The upper bound can be proven using the similar technique. Let P' be the distribution whose probability is concentrated at a maximum element, which means if $x = \arg\max_{x'} f(x')$, then, $P'(x) = 1$ and, otherwise, $P'(x) = 0$. If there are multiple maximum at $f(x)$, then, one of them can be arbitrarily chosen. Then, the Tsallis entropy of P' becomes zero since all probability mass is concentrated at a single instance, i.e., $S_q(P') = 0$. Then, the upper bound can be computed as follows:

$$\begin{aligned} q\text{-}\max_x(f(x)) &= \max_{P\in\Delta}\left[\mathop{\mathbb{E}}_{X\sim P}[f(X)] + S_q(P)\right] \\ &\ge \mathop{\mathbb{E}}_{X\sim P'}[f(X)] + S_q(P') = f\left(\arg\max_{x'} f(x')\right) = \max_x f(x). \end{aligned} \tag{5}$$

■

We now analyze the solution of q-maximum operator.

Proposition 3. *The optimal solution of (2) is*

$$\pi_q^\star(a) = \exp_q\left(\frac{\mathbf{r}(a)}{q} - \psi_q\left(\frac{\mathbf{r}}{q}\right)\right), \tag{6}$$

where the q-potential function ψ_q is a functional defined on $\{\mathcal{A}, \mathbf{r}\}$. ψ_q is determined uniquely for given $\{\mathcal{A}, \mathbf{r}\}$ by the following normalization condition:

$$\sum_a \pi_q^\star(a) = \sum_a \exp_q\left(\frac{\mathbf{r}(a)}{q} - \psi_q\left(\frac{\mathbf{r}}{q}\right)\right) = 1. \tag{7}$$

Furthermore, using $\pi_q^\star$, the optimal value can be written as

$$\mathop{\mathbb{E}}_{a\sim\pi^\star}[R] + S_q(\pi^\star) = (q-1)\sum_a \frac{\mathbf{r}(a)}{q}\exp_q\left(\frac{\mathbf{r}(a)}{q} - \psi_q\left(\frac{\mathbf{r}}{q}\right)\right) + \psi_q\left(\frac{\mathbf{r}}{q}\right). \tag{8}$$

Proof: It is easy to check ψ_q exists uniquely for given $\{\mathcal{A}, \mathbf{r}\}$. Indeed, because $\exp_q \in [0, \infty)$ is a continuous monotonic function, for any $\{\mathcal{A}, \mathbf{r}\}$, $\sum_a \exp_q\left(\frac{\mathbf{r}}{q} - \xi\right)$ converge to 0 and ∞ if ξ goes to $+\infty$ and $-\infty$, respectively. Therefore by the intermediate value theorem, there exists a unique constant $\xi^* \in \mathbb{R}$ such that $\sum_a \exp_q\left(\frac{\mathbf{r}(a)}{q} - \xi^*\right) = 1$. Hence it is sufficient to take $\psi_q(\mathbf{r}/q) = \xi^*$.

To show the remains, we mainly employ the convex optimization technique. Since $S_q(\pi)$ is concave and the expectation and constraints of Δ are linear w. r. t. π, the problem is concave. Thus, strong duality holds and we can use KKT conditions to obtain an optimal solution.

Δ has two constraints: sum-to-one and nonnegativity. Let μ be a dual variable for $1 - \sum_a \pi(a) = 0$ and $\lambda(a)$ be a dual variable for $\pi(a) \geq 0$. Then, KKT conditions are as follows:

$$\begin{aligned}
&\forall\, i \;\; 1 - \sum_a \pi_q^\star(a) = 0, \;\; \pi_q^\star(a) \geq 0 \\
&\forall\, i \;\; \lambda^\star(a) \leq 0 \\
&\forall\, i \;\; \lambda^\star(a) p_i^\star = 0 \\
&\forall\, i \;\; \mathbf{r}(a) - \mu^\star - \ln_q(\pi_q^\star(a)) - (\pi_q^\star(a))^{q-1} + \lambda^\star(a) = 0
\end{aligned} \tag{9}$$

where * indicates an optimal solution. We focus on the last condition. The last condition is converted into

$$\begin{aligned}
0 &= \mathbf{r}(a) - \mu^\star - \ln_q(\pi_q^\star(a)) - (\pi_q^\star(a))^{q-1} + \lambda^\star(a) \\
0 &= \mathbf{r}(a) - \mu^\star - \ln_q(\pi_q^\star(a)) - (q-1)\frac{\pi_q^\star(a)^{q-1} - 1}{q-1} - 1 + \lambda^\star(a) \\
0 &= \mathbf{r}(a) - \mu^\star - q\ln_q(\pi_q^\star(a)) - 1 + \lambda^\star(a)
\end{aligned} \tag{10}$$

First, let's consider positive measure $\pi_q^\star(a) > 0$ ($\lambda^\star(a) = 0$). Then, from equation (10),

$$\exp_q\left(\frac{\mathbf{r}(a)}{q} - \frac{\mu^\star + 1}{q}\right) = \pi_q^\star(a) \tag{11}$$

and $\mu^\star$ can be found by solving the following equation:

$$\sum_a \exp_q\left(\frac{\mathbf{r}(a)}{q} - \frac{\mu^\star + 1}{q}\right) = 1. \tag{12}$$

Since the equation (12) is exactly same as a normalization equation (7), $\mu^\star$ can be found using a q-potential function ψ_q:

$$\mu^\star = q\psi_q\left(\frac{\mathbf{r}}{q}\right) - 1 \tag{13}$$

Then,

$$\pi_q^\star(a) = \exp_q\left(\frac{\mathbf{r}(a)}{q} - \psi_q\left(\frac{\mathbf{r}}{q}\right)\right). \tag{14}$$

The optimal value of primal problem is

$$\begin{aligned}
&\mathop{\mathbb{E}}_{a\sim\pi_q^\star}[R] + S_q(\pi_q^\star) = \sum_a \mathbf{r}(a)\pi_q^\star(a) - \sum_a \left[\frac{\mathbf{r}(a)}{q} - \psi_q\left(\frac{\mathbf{r}}{q}\right)\right]\pi_q^\star(a) \\
&= (q-1)\sum_a \frac{\mathbf{r}(a)}{q}\exp_q\left(\frac{\mathbf{r}(a)}{q} - \psi_q\left(\frac{\mathbf{r}}{q}\right)\right) + \psi_q\left(\frac{\mathbf{r}}{q}\right).
\end{aligned} \tag{15}$$

Finally, let's check the supporting set. For $\pi_q^\star(a) > 0$, the following condition should be satisfied:

$$1 + (q-1)\left(\frac{\mathbf{r}(a)}{q} - \psi_q\left(\frac{\mathbf{r}}{q}\right)\right) > 0, \tag{16}$$

where this condition comes from the definition of $\exp_q(x)$. ■

Markov Decision Processes with Tsallis entropy maximization is formulated as follows.

$$\mathop{\text{maximize}}_{\pi\in\Pi} \quad \mathop{\mathbb{E}}_{\tau\sim P,\pi}\left[\sum_t^\infty \gamma^t \mathbf{R}_t\right] + \alpha S_q^\infty(\pi) \tag{17}$$

In this section, we analyze the optimality condition of a Tsallis MDP.

Theorem 2. *An optimal policy $\pi_q^\star$ and optimal value $V_q^\star$ sufficiently and necessarily satisfy the following Tsallis-Bellman optimality (TBO) equations:*

$$\begin{aligned}
Q_q^\star(s,a) &= \mathop{\mathbb{E}}_{s'\sim P}[\mathbf{r}(s,a,s') + \gamma V_q^\star(s')|s,a] \\
V_q^\star(s) &= q\text{-}\max_a(Q_q^\star(s,a)) \\
\pi_q^\star(a|s) &= \exp_q\left(\frac{Q_q^\star(s,a)}{q} - \psi_q\left(\frac{Q_q^\star(s,\cdot)}{q}\right)\right),
\end{aligned} \tag{18}$$

where ψ_q is a q-potential function.

Before starting proof, we first remind two propositions and prove one lemma. They are mainly employed to convert the optimization variable from π to the state action visitation ρ.

Proposition 4. *Let a state visitation be $\rho_\pi(s) = \mathbb{E}_{\tau\sim P,\pi}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{1}(s_t = s)\right]$ and state action visitation be $\rho_\pi(s,a) = \mathbb{E}_{\tau\sim P,\pi}\left[\sum_{t=0}^{\infty}\gamma^t \mathbb{1}(s_t = s, a_t = a)\right]$. Following relationships hold.*

$$\rho_\pi(s) = \sum_a \rho_\pi(s,a), \quad \rho_\pi(s,a) = \rho_\pi(s)\pi(a|s) \tag{19}$$

$$\sum_a \rho_\pi(s,a) = d(s) + \sum_{s',a'} P(s|s',a')\rho_\pi(s',a'), \quad \rho_\pi(s,a) \tag{20}$$

where Equation (20) is called Bellman Flow constraints.

Proof: Proof can be found in [2, 3] ■

Proposition 4 tells us, for fixed policy π, ρ_π satisfies Bellman Flow constraints. Then, the next remark show the opposite direction where if some function ρ satisfies Bellman Flow constraints, then there exist an unique policy which induces ρ.

Proposition 5 (Theorem 2 of Syed et al. [3])**.** *Let $\mathbf{M}$ be a set of state-action visitation measures, i.e., $\mathbf{M} \triangleq \{\rho | \forall s,\ a,\ \rho(s,a) \geq 0,\ \sum_a \rho(s,a) = d(s) + \sum_{s',a'} P(s|s',a')\rho(s',a')\}$. If $\rho \in \mathbf{M}$, then it is a state-action visitation measure for $\pi_\rho(a|s) \triangleq \frac{\rho(s,a)}{\sum_{a'}\rho(s,a')}$, and π_ρ is the unique policy whose state-action visitation measure is ρ.*

Proof: Proof can be found in [2, 3]. ■

Now, proposition 4 and 5 tell us that a policy and state action visitation have the one-to-one correspondence. In the following lemmas, we convert the optimization variable from π to ρ based on one-to-one correspondence.

Lemma 2.1. *Let*

$$\bar{S}_q^\infty(\rho) = -\sum_{s,a}\rho(s,a)\ln_q\left(\frac{\rho(s,a)}{\sum_{a'}\rho(s,a')}\right).$$

Then, for any stationary policy $\pi \in \Pi$ and any state-action visitation measure $\rho \in \mathbf{M}$, $S_q^\infty(\pi) = \bar{S}_q^\infty(\rho_\pi)$ and $\bar{S}_q^\infty(\rho) = S_q^\infty(\pi_\rho)$ hold.

Proof: First, show that $S_q^\infty(\pi) = \bar{S}_q^\infty(\rho_\pi)$.

$$\begin{aligned} S_q^\infty(\pi) &= \mathop{\mathbb{E}}_{\tau\sim P,\pi}\left[\sum_{t=0}^{\infty}\gamma^t S_q(\pi(\cdot|s_t))\right] = \sum_s S_q(\pi(\cdot|s))\cdot\mathbb{E}_{\tau\sim P,\pi}\left[\sum_{t=0}^{\infty}\gamma^t\mathbb{1}(s_t=s)\right] \\ &= \sum_s S_q(\pi(\cdot|s))\rho_\pi(s) = \sum_{s,a} -\ln_q(\pi(a|s))\pi(a|s)\rho_\pi(s) \\ &= \sum_{s,a} -\ln_q\left(\frac{\rho_\pi(s,a)}{\sum_{a'}\rho_\pi(s,a')}\right)\rho_\pi(s,a) = \bar{S}_q^\infty(\rho_\pi) \end{aligned} \tag{21}$$

Next, show that $\bar{S}_q^\infty(\rho) = S_q^\infty(\pi_\rho)$.

$$\bar{S}_q^\infty(\rho) = \sum_{s,a} -\ln_q\left(\frac{\rho(s,a)}{\sum_{a'}\rho(s,a')}\right)\rho(s,a) = \sum_{s,a} -\ln_q(\pi_\rho(a|s))\pi_\rho(a|s)\rho(s) = S_q^\infty(\pi_\rho) \tag{22}$$

■

Corollary 2.1.1. *The problem (23) is equivalent to a Tsallis MDP, which means if $\rho^\star$ is an optimal solution of (23), then, $\pi_{\rho^\star}$ is an optimal solution of a Tsallis MDP and vice versa.*

Proof: Let $\rho^\star$ be an optimal solution of (23). Assume that there exist another policyt π' such that $J(\pi') + S_q^\infty(\pi') > J(\pi_{\rho^\star}) + S_q^\infty(\pi_{\rho^\star})$ where $J(\pi) = \mathbb{E}_{\tau\sim\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t\mathbf{R}_t\right]$. Then, $\sum_{s,a}\rho_{\pi'}(s,a)\mathbf{r}(s,a) + \bar{S}_q^\infty(\rho_{\pi'}) > \sum_{s,a}\rho_\star(s,a)\mathbf{r}(s,a) + \bar{S}_q^\infty(\rho_\star)$. It contradicts to the fact that $\rho^\star$ is the optimal solution of (23). Thus, for all π, $J(\pi) + S_q^\infty(\pi) \leq J(\pi_{\rho^\star}) + S_q^\infty(\pi_{\rho^\star})$ which means $\pi_{\rho^\star}$ is the optimal policy. The opposite direction also can be proven in the same way. ■

Lemma 2.1 shows that $\bar{S}_q^\infty(\rho)$ and $S_q^\infty(\pi)$ has the same function value. Thus, we can freely change the optimization variable from π to ρ since the optimal point does not change due to the Corollary 2.1.1.

A. Variable Change

Based on Proposition 5 and Lemma 2.1, we convert a Tsallis MDP problem to

$$\begin{aligned} \underset{\rho}{\text{maximize}} \quad & \sum_{s,a} \rho(s,a) \sum_{s'} \mathbf{r}(s,a,s') P(s'|s,a) - \sum_{s,a} \rho(s,a) \ln_q \left(\frac{\rho(s,a)}{\sum_{a'} \rho(s,a')} \right) \\ \text{subject to} \quad & \forall s,\ a,\ \rho(s,a) \geq 0,\ \sum_{a} \rho(s,a) = d(s) + \sum_{s',a'} P(s|s',a') \rho(s',a'). \end{aligned} \tag{23}$$

Now, the optimization variables in the problem (23) is a state action visitation. In the following lemmas, we show that the problem (23) is concave with respect to a state action visitation.

Lemma 2.2. *$\bar{S}_q^{\infty}(\rho)$ is concave function with respect to $\rho \in \mathbf{M}$*

Proof: Let us consider the function $f(x) = -x \ln_q(x)$ defined over $(x > 0)$. Second derivative of $d^2 f(x)/dx^2$ is computed as

$$\frac{d^2 f(x)}{dx^2} = -q x^{q-2} < 0 \ \ (x > 0).$$

Since its second derivative is always negative on its domain, $f(x)$ is a concave function. From this fact, we can show that $\bar{S}_q^{\infty}(\rho)$ is concave. Proving the concavity is equivalent to show that for any $0 < \lambda_1, \lambda_2 < 1$ such that $\lambda_1 + \lambda_2 = 1$, and for $\rho_1, \rho_2 \in \mathbf{M}$ the following inequality holds

$$\bar{S}_q^{\infty}(\lambda_1 \rho_1 + \lambda_2 \rho_2) > \lambda_1 \bar{S}_q^{\infty}(\rho_1) + \lambda_2 \bar{S}_q^{\infty}(\rho_2)$$

For notional simplicity, let $\tilde{\rho}$ be $\lambda_1 \rho_1 + \lambda_2 \rho_2$ and define $\mu_1 = \frac{\lambda_1 \sum_{a'} \rho_1(s,a')}{\sum_{a'} \tilde{\rho}(s,a')}$ and $\mu_2 = \frac{\lambda_2 \sum_{a'} \rho_2(s,a')}{\sum_{a'} \tilde{\rho}(s,a')}$. Note that from the definition, $\mu_1 + \mu_2 = 1$. It can be shown as follow:

$$\begin{aligned} \bar{S}_q^{\infty}(\lambda_1 \rho_1 + \lambda_2 \rho_2) &= -\sum_{s,a} \tilde{\rho}(s,a) \ln_q \left(\frac{\lambda_1 \rho_1(s,a) + \lambda_2 \rho_2(s,a)}{\sum_{a'} \tilde{\rho}(s,a')} \right) \\ &= -\sum_{s,a} \tilde{\rho}(s,a) \ln_q \left(\frac{\mu_1 \rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} + \frac{\mu_2 \rho_2(s,a)}{\sum_{a'} \rho_2(s,a')} \right) \\ &= -\sum_{s,a} \left(\sum_{a'} \tilde{\rho}(s,a') \frac{\lambda_1 \rho_1(s,a) + \lambda_2 \rho_2(s,a)}{\sum_{a'} \tilde{\rho}(s,a')} \right) \ln_q \left(\frac{\mu_1 \rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} + \frac{\mu_2 \rho_2(s,a)}{\sum_{a'} \rho_2(s,a')} \right) \\ &= -\sum_{s,a} \sum_{a'} \tilde{\rho}(s,a') \left(\frac{\mu_1 \rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} + \frac{\mu_2 \rho_2(s,a)}{\sum_{a'} \tilde{\rho}(s,a')} \right) \ln_q \left(\frac{\mu_1 \rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} + \frac{\mu_2 \rho_2(s,a)}{\sum_{a'} \rho_2(s,a')} \right) \end{aligned} \tag{24}$$

Then, for all s, a,

$$\begin{aligned} & -\left(\frac{\mu_1 \rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} + \frac{\mu_2 \rho_2(s,a)}{\sum_{a'} \tilde{\rho}(s,a')} \right) \ln_q \left(\frac{\mu_1 \rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} + \frac{\mu_2 \rho_2(s,a)}{\sum_{a'} \rho_2(s,a')} \right) \\ & > -\mu_1 \frac{\rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} \ln_q \left(\frac{\rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} \right) - \mu_2 \frac{\rho_2(s,a)}{\sum_{a'} \rho_2(s,a')} \ln_q \left(\frac{\rho_2(s,a)}{\sum_{a'} \rho_2(s,a')} \right) \end{aligned}$$

Equation (24) becomes

$$\begin{aligned} & \bar{S}_q^{\infty}(\lambda_1 \rho_1 + \lambda_2 \rho_2) \\ & = -\sum_{s,a} \sum_{a'} \tilde{\rho}(s,a') \left(\frac{\mu_1 \rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} + \frac{\mu_2 \rho_2(s,a)}{\sum_{a'} \tilde{\rho}(s,a')} \right) \ln_q \left(\frac{\mu_1 \rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} + \frac{\mu_2 \rho_2(s,a)}{\sum_{a'} \rho_2(s,a')} \right) \\ & > -\sum_{s,a} \sum_{a'} \tilde{\rho}(s,a') \mu_1 \frac{\rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} \ln_q \left(\frac{\rho_1(s,a)}{\sum_{a'} \rho_1(s,a')} \right) \\ & \quad -\sum_{s,a} \sum_{a'} \tilde{\rho}(s,a') \mu_2 \frac{\rho_2(s,a)}{\sum_{a'} \rho_2(s,a')} \ln_q \left(\frac{\rho_2(s,a)}{\sum_{a'} \rho_2(s,a')} \right) \end{aligned}$$

Since $\sum_{a'} \tilde{\rho}(s,a')\mu_1 \frac{\rho_1(s,a)}{\sum_{a'}\rho_1(s,a')} = \sum_{a'} \tilde{\rho}(s,a') \frac{\lambda_1 \sum_{a'}\rho_1(s,a')}{\sum_{a'}\tilde{\rho}(s,a')} \frac{\rho_1(s,a)}{\sum_{a'}\rho_1(s,a')} = \lambda_1 \rho_1(s,a)$, finally, we get

$$
\begin{aligned}
&\bar{S}_q^{\infty}(\lambda_1\rho_1 + \lambda_2\rho_2) \\
&> -\sum_{s,a}\sum_{a'} \tilde{\rho}(s,a')\mu_1 \frac{\rho_1(s,a)}{\sum_{a'}\rho_1(s,a')} \ln_q\left(\frac{\rho_1(s,a)}{\sum_{a'}\rho_1(s,a')}\right) \\
&\quad -\sum_{s,a}\sum_{a'} \tilde{\rho}(s,a')\mu_2 \frac{\rho_2(s,a)}{\sum_{a'}\rho_2(s,a')} \ln_q\left(\frac{\rho_2(s,a)}{\sum_{a'}\rho_2(s,a')}\right) \\
&= -\sum_{s,a} \lambda_1\rho_1(s,a) \ln_q\left(\frac{\rho_1(s,a)}{\sum_{a'}\rho_1(s,a')}\right) - \sum_{s,a} \lambda_2\rho_2(s,a) \ln_q\left(\frac{\rho_2(s,a)}{\sum_{a'}\rho_2(s,a')}\right) \\
&= \lambda_1 \bar{S}_q^{\infty}(\rho_1) + \lambda_2 \bar{S}_q^{\infty}(\rho_2)
\end{aligned}
$$

Note that this proof holds for every q value greater than zero. ■

Corollary 2.2.1. *The problem (23) is concave with respect to* $\rho \in \mathbf{M}$

Proof: The objective function of (23) is concave function w.r.t ρ since the first term is linear and the second term is concave be Lemma 2.2. All constraints are linear w.r.t ρ. Thus, the problem is a concave problem. ■

B. Proof of Tsallis Bellman Optimality Equation

Proof of Theorem 2: Since the problem (23) is concave with respect to ρ, the primal and dual solutions necessarily and sufficiently satisfy a KKT condition. First, the Lagrangian objecitve $\mathcal{L} \triangleq \sum_{s,a}\rho(s,a)\mathbf{r}(s,a) - \sum_{s,a}\rho(s,a)\ln_q\left(\frac{\rho(s,a)}{\sum_{a'}\rho(s,a')}\right) + \sum_{s,a}\lambda(s,a)\rho(s,a) + \sum_s \mu(s)\left(d(s) + \sum_{s',a'}P(s|s',a')\rho(s',a') - \sum_a \rho(s,a)\right)$ where $\lambda(s,a)$ and $\mu(s)$ are dual variables for nonnegativity and Bellman flow constraints. The KKT conditions of the problem (23) are as follows:

$$
\begin{aligned}
&\forall s,\ a,\ \rho^{\star}(s,a) \geq 0,\ d(s) + \sum_{s',a'} P(s|s',a')\rho^{\star}(s',a') - \sum_a \rho^{\star}(s,a) = 0 \\
&\forall s,\ a,\ \lambda^{\star}(s,a) \leq 0 \\
&\forall s,\ a,\ \lambda^{\star}(s,a)\rho^{\star}(s,a) = 0 \\
&\forall s,\ a,\ 0 = \sum_{s'} \mathbf{r}(s,a,s')P(s'|s,a) + \gamma \sum_{s'} \mu^{\star}(s')P(s'|s,a) \\
&\qquad - \mu^{\star}(s) - q\ln_q\left(\frac{\rho^{\star}(s,a)}{\sum_{a'}\rho^{\star}(s,a')}\right) - 1 + \sum_a \left(\frac{\rho^{\star}(s,a)}{\sum_{a'}\rho^{\star}(s,a')}\right)^q + \lambda^{\star}(s,a)
\end{aligned}
\tag{25}
$$

We would like to note that the dervative of $\bar{S}_q^{\infty}(\rho)$ is computed as follows:

$$
\begin{aligned}
\frac{\partial \bar{S}_q^{\infty}(\rho)}{\partial \rho(s'',a'')} &= -\sum_{s,a} \frac{\partial \rho(s,a)}{\partial \rho(s'',a'')} \ln_q\left(\frac{\rho(s,a)}{\sum_{a'}\rho(s,a')}\right) - \sum_{s,a}\rho(s,a)\frac{\partial \ln_q\left(\rho(s,a)/\sum_{a'}\rho(s,a')\right)}{\partial \rho(s'',a'')} \\
&= -\ln_q\left(\frac{\rho(s'',a'')}{\sum_{a'}\rho(s'',a')}\right) \\
&\quad -\sum_a \rho(s'',a)\left(\frac{\rho(s'',a)}{\sum_{a'}\rho(s'',a')}\right)^{q-2}\left(\frac{\delta_{a''}(a)}{\sum_{a'}\rho(s'',a')} - \frac{\rho(s'',a)}{\left(\sum_{a'}\rho(s'',a')\right)^2}\right) \\
&= -\ln_q\left(\frac{\rho(s'',a'')}{\sum_{a'}\rho(s'',a')}\right) - \left(\frac{\rho(s'',a'')}{\sum_{a'}\rho(s'',a')}\right)^{q-1} + \sum_a \left(\frac{\rho(s'',a)}{\sum_{a'}\rho(s'',a')}\right)^q \\
&= -q\ln_q\left(\frac{\rho(s'',a'')}{\sum_{a'}\rho(s'',a')}\right) - 1 + \sum_a \left(\frac{\rho(s'',a)}{\sum_{a'}\rho(s'',a')}\right)^q
\end{aligned}
\tag{26}
$$

Then, we show that $\mu^{\star}(s)$ is the same as optimal value $V_q^{\star}(s)$. From the stationary condition, by multiplying $\pi_{\rho^{\star}}(a|s) =$

$\rho^\star(s,a)/\sum_{a'}\rho^\star(s,a')$ and summing up with respect to a, the following equation is obtained:

$$\begin{aligned}
0 =& \sum_a \sum_{s'} \mathbf{r}(s,a,s') P(s'|s,a)\pi_{\rho^\star}(a|s) + \gamma \sum_{s'} \mu^\star(s') \sum_a P(s'|s,a)\pi_{\rho^\star}(a|s) - \mu^\star(s) \\
& - q \sum_a \pi_{\rho^\star}(a|s) \ln_q \left(\frac{\rho^\star(s,a)}{\sum_{a'} \rho^\star(s,a')} \right) - 1 + \sum_a \pi_{\rho^\star}(a|s) \sum_{a''} \left(\frac{\rho^\star(s,a'')}{\sum_{a'} \rho^\star(s,a')} \right)^q \\
& + \sum_a \lambda^\star(s,a) \pi_{\rho^\star}(a|s) \\
=& \sum_a \sum_{s'} \mathbf{r}(s,a,s') P(s'|s,a)\pi_{\rho^\star}(a|s) + \gamma \sum_{s'} \mu^\star(s') \sum_a P(s'|s,a)\pi_{\rho^\star}(a|s) \\
& - \mu^\star(s) - q \sum_a \pi_{\rho^\star}(a|s) \ln_q \left(\pi_{\rho^\star}(a|s) \right) - 1 + \sum_{a''} \pi_{\rho^\star}(s,a)^q + \sum_a \lambda^\star(s,a) \pi_{\rho^\star}(a|s) \\
=& \sum_a \sum_{s'} \mathbf{r}(s,a,s') P(s'|s,a)\pi_{\rho^\star}(a|s) + \gamma \sum_{s'} \mu^\star(s') \sum_a P(s'|s,a)\pi_{\rho^\star}(a|s) \\
& - \mu^\star(s) - q \sum_a \pi_{\rho^\star}(a|s) \ln_q \left(\pi_{\rho^\star}(a|s) \right) - 1 + \sum_{a''} \pi_{\rho^\star}(s,a)^q \\
=& \sum_a \sum_{s'} \mathbf{r}(s,a,s') P(s'|s,a)\pi_{\rho^\star}(a|s) + \gamma \sum_{s'} \mu^\star(s') \sum_a P(s'|s,a)\pi_{\rho^\star}(a|s) \\
& - \mu^\star(s) - q \sum_a \pi_{\rho^\star}(a|s) \ln_q \left(\pi_{\rho^\star}(a|s) \right) + (q-1) \sum_{s,a} \pi_{\rho^\star}(s,a) \ln_q \left(\pi_{\rho^\star}(s,a) \right) \\
=& \sum_a \sum_{s'} \mathbf{r}(s,a,s') P(s'|s,a)\pi_{\rho^\star}(a|s) + \gamma \sum_{s'} \mu^\star(s') \sum_a P(s'|s,a)\pi_{\rho^\star}(a|s) \\
& - \mu^\star(s) - \sum_{s,a} \pi_{\rho^\star}(s,a) \ln_q \left(\pi_{\rho^\star}(s,a) \right).
\end{aligned} \tag{27}$$

Finally,

$$\begin{aligned}
\mu^\star(s) =& \sum_a \sum_{s'} \mathbf{r}(s,a,s') P(s'|s,a)\pi_{\rho^\star}(a|s) + \gamma \sum_{s'} \mu^\star(s') \sum_a P(s'|s,a)\pi_{\rho^\star}(a|s) \\
& - \sum_a \pi_{\rho^\star}(a|s) \ln_q \left(\pi_{\rho^\star}(a|s) \right) \\
=& \mathbb{E}_{s' \sim P, a \sim \pi} \left[\mathbf{r}(s,a,s') + \alpha S_q(\pi_{\rho^\star}(\cdot|s)) + \gamma \mu^\star(s') | s \right]
\end{aligned} \tag{28}$$

This equation (28) exactly satisfies Tsallis Bellman expectation (TBE) equation of $\pi_{\rho^\star}$. Thus, we want to claim that $\mu^\star(s)$ is the value $V^{\pi_{\rho^\star}}(s)$ of optimal policy $\pi_{\rho^\star}$, i.e., $\mu^\star(s) = V_q^\star(s)$. However, to guarantee $\mu^\star(s) = V_q^\star(s)$, we should prove the following statement: *if an arbitrary function $f(s)$ satisfies a TBE equation for π, then, $f(s) = V^\pi(s)$*, which is not yet proven but it will be in Theorem 3. In this proof, let us just believe Theorem 3.

Then, we first analyze a positive state-action visitation $\rho^\star(s,a) > 0$ ($\lambda^\star(s,a) = 0$). Using the fact that $\mu^\star = V_q^\star$, we can obtain $Q_q^\star(s,a) = \mathbb{E}_{s' \sim P}[\mathbf{r}(s,a,s') + \gamma \mu^\star(s')]$. By replacing $\rho^\star(s,a)/\sum_{a'}\rho^\star(s,a')$ with $\pi_{\rho^\star}(a|s)$ and using $Q_q^\star(s,a) = \mathbb{E}_{s' \sim P}[\mathbf{r}(s,a,s') + \gamma \mu^\star(s')]$ and $\mu^\star(s) = V^\star(s)$,

$$\begin{aligned}
& Q_q^\star(s,a) - V_q^\star(s) - q \ln_q \left(\pi_{\rho^\star}(a|s) \right) - 1 + \sum_a \pi_{\rho^\star}(a|s)^q = 0 \\
& \frac{Q_q^\star(s,a)}{q} - \frac{V_q^\star(s) + 1 - \sum_a \left(\pi_{\rho^\star}(a|s) \right)^q}{q} = \ln_q \left(\pi_{\rho^\star}(a|s) \right) \\
& \exp_q \left(\frac{Q_q^\star(s,a)}{q} - \frac{V_q^\star(s) + 1 - \sum_a \left(\pi_{\rho^\star}(a|s) \right)^q}{q} \right) = \pi_{\rho^\star}(a|s).
\end{aligned} \tag{29}$$

Now, we can use $\sum_a \pi(a|s) = 1$. By summing up with respect to a,

$$\sum_a \exp_q \left(\frac{Q_q^\star(s,a)}{q} - \frac{V_q^\star(s) + 1 - \sum_a \left(\pi_{\rho^\star}(a|s) \right)^q}{q} \right) = 1. \tag{30}$$

This equation is the normalization equation of q-exponential distribution (7). So, we can obtain the relationship between q-potential and the optimal value function.

$$\psi_q \left(\frac{Q_q^\star(s,\cdot)}{q} \right) = \frac{V_q^\star(s) + 1 - \sum_a \left(\pi_{\rho^\star}(a|s) \right)^q}{q} \tag{31}$$

Finally, it is shown that the optimal policy has q-exponential distribution of $Q_q^\star(s,\cdot)$.

$$\exp_q\left(\frac{Q_q^\star(s,a)}{q}-\psi_q\left(\frac{Q_q^\star(s,\cdot)}{q}\right)\right)=\pi_{\rho^\star}(a|s) \tag{32}$$

By plugging in this result into (28),

$$\begin{aligned}
V_q^\star(s)&=\sum_a \pi_{\rho^\star}(a|s)\sum_{s'}\left[\mathbf{r}(s,a,s')+\gamma V_q^\star(s')P(s'|s,a)\right]-\sum_a \pi_{\rho^\star}(a|s)\ln_q\left(\pi_q^\star(a|s)\right)\\
&=\sum_a \pi_{\rho^\star}(a|s)Q_q^\star(s,a)-\sum_a \pi_{\rho^\star}(a|s)\ln_q\left(\pi_q^\star(a|s)\right)\\
&=\sum_a \pi_{\rho^\star}(a|s)Q_q^\star(s,a)-\sum_a \pi_{\rho^\star}(a|s)\left(\frac{Q_q^\star(s,a)}{q}-\psi_q\left(\frac{Q_q^\star(s,\cdot)}{q}\right)\right)\\
&=(q-1)\sum_a \pi_{\rho^\star}(a|s)\frac{Q_q^\star(s,a)}{q}+\psi_q\left(\frac{Q_q^\star(s,\cdot)}{q}\right)\\
&=q\text{-}\max_{a'}\left(Q_q^\star(s,a')\right)
\end{aligned} \tag{33}$$

where the last equation is derived using the Equation (8).

To summarize, we obtain the optimality condition for a Tsallis MDP as follows:

$$\begin{aligned}
Q_q^\star(s,a)&=\mathbb{E}_{s'}\left[\mathbf{r}(s,a,s')+\gamma V^\star(s')|s,a\right]\\
V_q^\star(s)&=q\text{-}\max_{a'}(Q_q^\star(s,a'))\\
\pi_q^\star(a|s)&=\exp_q\left(\frac{Q_q^\star(s,a)}{q}-\psi_q\left(\frac{Q_q^\star(s,\cdot)}{q}\right)\right)
\end{aligned} \tag{34}$$

We call these equations Tsallis Bellman optimality (TBO) equations. ■

C. Tsallis Bellman Expectation (TBE) Equation

In Tsallis policy evaluation, for fixed π, the value functions of π have the relationship as follows:

$$\begin{aligned}
Q_q^\pi(s,a)&=\mathop{\mathbb{E}}_{s'\sim P}[\mathbf{r}(s,a,s')+\gamma V_q^\pi(s')|s,a]\\
V_q^\pi(s)&=\mathop{\mathbb{E}}_{a\sim\pi}[Q_q^\pi(s,a)-\ln_q(\pi(a|s))],
\end{aligned} \tag{35}$$

These equations are derived from the definition of V_q^π and Q_q^π. Thus, if we have some value functions of Tsallis MDP, then, they satisfies TBE equation trivially. However, main goal of Tsallis policy evaluation is to prove the opposite direction: *if an arbitrary function $f(s)$ satisfies a TBE equation for π, then, $f(s)=V^\pi(s)$.*

D. Tsallis Bellman Expectation Operator and Tsallis Policy Evaluation

$$\begin{aligned}
\left[\mathcal{T}_q^\pi F\right](s,a)&\triangleq\mathop{\mathbb{E}}_{s'\sim P}[\mathbf{r}(s,a,s')+\gamma V_F(s')|s,a]\\
V_F(s)&\triangleq\mathop{\mathbb{E}}_{a\sim\pi}[F(s,a)-\ln_q(\pi(a|s))],
\end{aligned} \tag{36}$$

where $s'\sim P$ indicates $s'\sim P(\cdot|s,a)$ and $a'\sim\pi$ indicates $a'\sim\pi(\cdot|s)$. Then, policy evaluation method in a Tsallis MDP can be simply defined as

$$F_{k+1}=\mathcal{T}_q^\pi F_k.$$

Theorem 3 (Tsallis Policy Evaluation)**.** *For any fixed policy π and entropic-index $q\geq 1$, consider Tsallis Bellman expectation (TBE) operator $\mathcal{T}_q^\pi$, and for an arbitrary initial function F over $\mathcal{S}\times\mathcal{A}$, define Tsallis policy evaluation $F_{k+1}=\mathcal{T}_q^\pi F_k$. Then, F_k converges into the Q_q^π and satisfies TBE equation (35). In other words, the value function satisfying the TBE equation (35) is unique.*

Before proving Theorem 3, we first drive the properties of $\mathcal{T}_q^\pi$.

Lemma 3.1. *For $F:\mathcal{S}\times\mathcal{A}\to R$ and $c\in R^+$, $\mathcal{T}_q^\pi(F+c\mathbf{1})=\mathcal{T}_q^\pi F+\gamma c\mathbf{1}$ where $\mathbf{1}:\mathcal{S}\times\mathcal{A}\to 1$*

Proof: For all s,a,

$$\begin{aligned}
V_{F+c\mathbf{1}}(s)&=\mathop{\mathbb{E}}_{a\sim\pi}[F(s,a)+c-\ln_q(\pi(a|s))]=\mathop{\mathbb{E}}_{a\sim\pi}[F(s,a)-\ln_q(\pi(a|s))]+c\\
&=V_F(s)+c
\end{aligned} \tag{37}$$

Thus,

$$\begin{aligned}\left[\mathcal{T}_q^\pi\left(F+c\mathbf{1}\right)\right](s,a) &= \mathop{\mathbb{E}}_{s'\sim P}[\mathbf{r}(s,a,s')+\gamma V_{F+c\mathbf{1}}(s')|s,a] \\ &= \mathop{\mathbb{E}}_{s'\sim P}[\mathbf{r}(s,a,s')+\gamma V_F(s')+\gamma c|s,a] \\ &= \mathop{\mathbb{E}}_{s'\sim P}[\mathbf{r}(s,a,s')+\gamma V_F(s')|s,a]+\gamma c = \mathcal{T}_q^\pi F(s)+\gamma c\end{aligned} \tag{38}$$

■

Lemma 3.2. *For $F, G : \mathcal{S}\times\mathcal{A}\to R$ and $F \succeq G$, $\mathcal{T}_q^\pi(F) \succeq \mathcal{T}_q^\pi(G)$ where $\succeq$ indicates $F(s,a)\geq G(s,a)$ for all s,a.*

Proof: For all s, a,

$$V_F(s) = \mathop{\mathbb{E}}_{a\sim\pi}[F(s,a)-\ln_q(\pi(a|s))] < \mathop{\mathbb{E}}_{a\sim\pi}[G(s,a)-\ln_q(\pi(a|s))] = V_G(s) \tag{39}$$

Thus,

$$\begin{aligned}\left[\mathcal{T}_q^\pi F\right](s,a) &= \mathop{\mathbb{E}}_{s'\sim P}[\mathbf{r}(s,a,s')+\gamma V_F(s')|s,a] \\ &< \mathop{\mathbb{E}}_{s'\sim P}[\mathbf{r}(s,a,s')+\gamma V_G(s')|s,a] = \left[\mathcal{T}_q^\pi G\right](s,a)\end{aligned} \tag{40}$$

■

Lemma 3.3. *$\mathcal{T}_q^\pi$ is γ-contraction mapping in $(C(\mathcal{S}\times\mathcal{A},R),|\cdot|_\infty)$ where $C(\mathcal{S}\times\mathcal{A},R)\triangleq\{F:\mathcal{S}\times\mathcal{A}\to R\}$ and $|F-G|_\infty = \sup_{s,a}|F(s,a)-G(s,a)|$*

Proof: Let $d = |F-G|_\infty$. The, $G - d\mathbf{1} \preceq F \preceq G + d\mathbf{1}$. From Lemma 3.2, $\mathcal{T}_q^\pi(G+d\mathbf{1}) \succeq \mathcal{T}_q^\pi F \succeq \mathcal{T}_q^\pi(G-d\mathbf{1})$. From Lemma 3.1, $\mathcal{T}_q^\pi G + \gamma d\mathbf{1} \succeq \mathcal{T}_q^\pi F \succeq \mathcal{T}_q^\pi G - \gamma d\mathbf{1}$. Then,$\gamma d\mathbf{1} \succeq \mathcal{T}_q^\pi F - \mathcal{T}_q^\pi G \succeq -\gamma d\mathbf{1}$. Finally,

$$|\mathcal{T}_q^\pi F - \mathcal{T}_q^\pi G|_\infty \leq \gamma d = \gamma|F-G|_\infty.$$

Consequently, $\mathcal{T}_q^\pi$ is γ-contraction. ■

1) Proof of Tsallis Policy Evaluation: *Proof of Theorem 3:* From Lemma 3.3, $\mathcal{T}_q^\pi$ is γ-contraction and has an unique fixed point $F_* = \mathcal{T}_q^\pi F_*$ from the Banach fixed point theorem. Then, for any initial function F, a sequence of F_k converges to the fixed point, i.e., $F_* = \lim_{k\to\infty}(\mathcal{T}_q^\pi)^k F_0$. The fixed point F_* satisfies a TBE equation as follows:

$$\begin{aligned}F_*(s,a) &= \mathop{\mathbb{E}}_{s'\sim P}[\mathbf{r}(s,a,s')+\gamma V_{F_*}(s')|s,a] \\ V_{F_*}(s) &= \mathop{\mathbb{E}}_{a\sim\pi}[F_*(s,a)-\ln_q(\pi(a|s))],\end{aligned} \tag{41}$$

Since F_* is unique, F_* is the only function which satisfies a TBE equation. Thus, $F_* = Q_q^\pi$. ■

E. Tsallis Policy Improvement

The value function evaluated from Tsallis policy evaluation can be employed to update the policy distribution. In policy improvement step, the policy will be updated to maximize

$$\forall s,\ \pi_{k+1}(\cdot|s) \triangleq \arg\max_{\pi(\cdot|s)} \mathop{\mathbb{E}}_{a\sim\pi}[Q_q^{\pi_k}(s,a)-\ln_q(\pi(a|s))|s] \tag{42}$$

Theorem 4 (Tsallis Policy Improvement)**.** *Let π_{k+1} be the updated policy from (42) using $Q_q^{\pi_k}$. For all $(s,a)\in\mathcal{S}\times\mathcal{A}$, $Q_q^{\pi_{k+1}}(s,a)$ is greater than $Q_q^{\pi_k}(s,a)$.*

Proof: , unless $\pi_k = \pi_{k+1}$ Since π_{k+1} is updated by maximizing Equation (42) and the maximization in Equation (42) is concave with respect to π, the following inequality holds

$$\mathop{\mathbb{E}}_{a\sim\pi_{k+1}}\left[Q_q^{\pi_k}(s,a)-\ln_q(\pi_{k+1}(a|s))\middle|s\right] \geq \mathop{\mathbb{E}}_{a\sim\pi_k}\left[Q_q^{\pi_k}(s,a)-\ln_q(\pi_k(a|s))\middle|s\right] = V_q^{\pi_k}(s), \tag{43}$$

where the equality holds when $\pi_{k+1} = \pi_k$. This inequality induces a performance improvement,

$$
\begin{aligned}
Q_q^{\pi_k}(s,a) &= \mathop{\mathbb{E}}_{s_1\sim P}\left[r(s_0,a_0,s_1) + \gamma V_q^{\pi_k}(s_1)\middle| s_0 = s, a_0 = a\right] \\
&\le \mathop{\mathbb{E}}_{s_1\sim P}\left[r(s_0,a_0,s_1)|s_0 = s, a_0 = a\right] \\
&\quad + \gamma \mathop{\mathbb{E}}_{s_1,a_1\sim P,\pi_{k+1}}\left[Q_q^{\pi_k}(s_1,a_1) - \ln_q(\pi_{k+1}(a_1|s_1))\middle| s_0 = s, a_0 = a\right] \\
&= \mathop{\mathbb{E}}_{s_1\sim P}\left[r(s_0,a_0,s_1)|s_0 = s, a_0 = a\right] \\
&\quad + \gamma \mathop{\mathbb{E}}_{s_{1:2},a_1\sim P,\pi_{k+1}}\left[r(s_1,a_1,s_2) - \ln_q(\pi_{k+1}(a_1|s_1)) + \gamma V_q^{\pi_k}(s_2)\middle| s_0 = s, a_0 = a\right] \\
&\le \mathop{\mathbb{E}}_{s_1\sim P}\left[r(s_0,a_0,s_1)|s_0 = s, a_0 = a\right] \\
&\quad + \gamma \mathop{\mathbb{E}}_{s_{1:t+1},a_{1:t}\sim P,\pi_{k+1}}\left[\sum_{k=1}^{t}\gamma^{k-1}\left(r(s_k,a_k,s_{k+1}) - \ln_q(\pi_{k+1}(a_k|s_k))\right)\middle| s_0 = s, a_0 = a\right] \\
&\quad + \gamma^{t+1}\mathop{\mathbb{E}}_{s_{t+1}\sim P,\pi_{k+1}}\left[V_q^{\pi_k}(s_{t+1})\middle| s_0 = s, a_0 = a\right] \\
&\vdots \\
&\le \mathop{\mathbb{E}}_{s_1\sim P}\left[r(s_0,a_0,s_1) + \gamma V_q^{\pi_{k+1}}(s_1)\middle| s_0 = s, a_0 = a\right] = Q_q^{\pi_{k+1}}(s,a),
\end{aligned}
\tag{44}
$$

where $\gamma^{t+1}\,\mathbb{E}_{s_{t+1}\sim P,\pi_{k+1}}\left[V_q^{\pi_k}(s_{t+1})\middle| s_0 = s, a_0 = a\right] \to 0$ as $t \to \infty$. ■

Theorem 5 (Optimality of TPI)**.** *TPI converges into an optimal policy and value of a Tsallis MDP.*

Proof: From the fact that reward function $\mathbf{r}$ has upper bound $\mathbf{r}_{\max}$ and $\mathcal{S}\times\mathcal{A}$ is bounded, $Q_q^{\pi_k}$ is also bouned. Then, since a sequence of $Q_q^{\pi_k}$ is monotonically non-decreasing and bounded, it converges to some point π_*. Now, proof will be done by showing $\pi_* = \pi_q^\star$. First, from the policy improvement, We have $\pi_*(\cdot|s) = \arg\max_{\pi(\cdot|s)}\mathbb{E}_{a\sim\pi}[Q_q^{\pi_*}(s,a) - \ln_q(\pi(a|s))|s]$ and at π_*, the equality in Equation (43) holds, i.e., $V_q^{\pi_*}(s) = \mathbb{E}_{a\sim\pi_*}\left[Q_q^{\pi_*}(s,a) - \alpha\ln_q(\pi_*(a|s))\middle|s\right]$. Then, the following equality holds,

$$
V_q^{\pi_*}(s) = \max_{\pi(\cdot|s)}\mathop{\mathbb{E}}_{a\sim\pi}\left[Q_q^{\pi_*}(s,a) - \alpha\ln_q(\pi(a|s))\middle|s\right],
$$

which is equivalent to $V_q^{\pi_*}(s) = q\text{-max}_{a'}\,Q^{\pi_*}(s,a')$. It can be also known that π_* is the solution of q-maximum. From the TBE equation, $Q_q^{\pi_*}(s,a) = \mathbb{E}_{s'\sim P}[\mathbf{r}(s,a,s') + \gamma V_q^{\pi_*}(s')|s,a]$. Thus, π_* satisfies a TBO equation and by Theorem 2, $\pi_* = \pi_q^\star$. ■

F. Tsallis Value Iteration

Tsallis value iteration is derived from the optimality equation. From TBO equation, Tsallis Bellman optimality operator is defined by

$$
\begin{aligned}
\left[\mathcal{T}_q F\right](s,a) &\triangleq \mathop{\mathbb{E}}_{s'\sim P}\left[\mathbf{r}(s,a,s') + \gamma V_F(s)|s,a\right] \\
V_F(s) &\triangleq \mathop{q\text{-max}}_{a'}\left(F(s,a')\right).
\end{aligned}
\tag{45}
$$

Then, a Tsallis value iteration is defined by repeatedly applying TBO operator:

$$
F_{k+1} = \mathcal{T}_q F_k.
$$

Theorem 6. *For any fixed entropic-index $q \ge 1$, consider Tsallis Bellman optimality (TBO) operator $\mathcal{T}_q$, and for an arbitrary initial function F_0 over $\mathcal{S}\times\mathcal{A}$, define Tsallis value iteration $F_{k+1} = \mathcal{T}_q F_k$. Then, F_k converges into the $Q_q^\star$.*

Before proving Theorem 6, we first drive the properties of q-maximum and $\mathcal{T}_q$.

Lemma 6.1. *For any function $f(x)$ defined on finite input space $\mathcal{X}$ and $c \in R$, The following equality hold:*

1) $q\text{-max}_x(f(x) + c\mathbf{1}) = q\text{-max}_x(f(x)) + c$
2) $q\text{-max}_x(f(x))$ *is monotone. If $f \preceq g$, then* $q\text{-max}_x(x) \le q\text{-max}_x(y)$

where $\mathbf{1}$ *is a constant function whose value is one.*

Proof: For property 1,

$$\begin{aligned} q\text{-}\max_x(f(x)+c\mathbf{1}) &= \max_{P\in\Delta}\left[\mathop{\mathbb{E}}_{X\sim P}[f(X)+c\mathbf{1}(X)]+S_q(P)\right] \\ &= \max_{P\in\Delta}\left[\mathop{\mathbb{E}}_{X\sim P}[f(X)]+c+S_q(P)\right] \\ &= \max_{P\in\Delta}\left[\mathop{\mathbb{E}}_{X\sim P}[f(X)]+S_q(P)\right]+c = q\text{-}\max_x(f(x))+c \end{aligned} \tag{46}$$

For property 2,

$$\begin{aligned} q\text{-}\max_x(f(x)) &= \max_{P\in\Delta}\left[\mathop{\mathbb{E}}_{X\sim P}[f(X)]+S_q(P)\right] \\ &= \mathop{\mathbb{E}}_{X\sim P^\star(f)}[f(X)]+S_q(P^\star(f)) \le \mathop{\mathbb{E}}_{X\sim P^\star(f)}[g(X)]+S_q(P^\star(f)) \quad (\because f\preceq g) \\ &\le \max_{P'\in\Delta}\left[\mathop{\mathbb{E}}_{X\sim P'}[g(X)]+S_q(P')\right] = q\text{-}\max_x(f(x)), \end{aligned} \tag{47}$$

where $P^\star(f)$ indicates the optimal distribution of $q\text{-}\max_x(f(x))$. ■

Lemma 6.2. *For $F:\mathcal{S}\times\mathcal{A}\to R$ and $c\in R$, $\mathcal{T}_q\left(F+c\mathbf{1}\right)=\mathcal{T}_qF+\gamma c\mathbf{1}$ where $\mathbf{1}:\mathcal{S}\times\mathcal{A}\to 1$*

Proof: For all s,a,

$$\begin{aligned} V_{F+c\mathbf{1}}(s) &= \mathop{q\text{-}\max}_{a'}\left(F(s,a')+c\right) = \mathop{q\text{-}\max}_{a'}\left(F(s,a')\right)+c = V_F(s)+c \\ \left[\mathcal{T}_qF+c\mathbf{1}\right](s,a) &= \mathop{\mathbb{E}}_{s'\sim P}[r(s,a,s')+\gamma V_{F+c\mathbf{1}}(s')|s,a] \\ &= \mathop{\mathbb{E}}_{s'\sim P}[r(s,a,s')+\gamma V_F(s')+\gamma c|s,a] \\ &= \mathop{\mathbb{E}}_{s'\sim P}[r(s,a,s')+\gamma V_F(s')|s,a]+\gamma c = \left[\mathcal{T}_qF\right](s,a)+\gamma c \end{aligned} \tag{48}$$

■

Lemma 6.3. *For $F,G:\mathcal{S}\times\mathcal{A}\to R$ and $F\succeq G$, $\mathcal{T}_q\left(F\right)\succeq\mathcal{T}_q\left(G\right)$ where $\succeq$ indicates $F(s,a)\ge G(s,a)$ for all s,a.*

Proof: For all s,a,

$$\begin{aligned} V_F(s) &= \mathop{q\text{-}\max}_{a'}\left(F(s,a')\right) \le \mathop{q\text{-}\max}_{a'}\left(G(s,a')\right) = V_G(s) \\ \left[\mathcal{T}_qF\right](s,a) &= \mathop{\mathbb{E}}_{s'\sim P}[r(s,a,s')+\gamma V_F(s')|s,a] \\ &\le \mathop{\mathbb{E}}_{s'\sim P}[r(s,a,s')+\gamma V_G(s')|s,a] = \left[\mathcal{T}_qG\right](s,a) \end{aligned} \tag{49}$$

■

Lemma 6.4. *$\mathcal{T}_q$ is γ-contraction mapping in $(C(\mathcal{S}\times\mathcal{A},R),|\cdot|_\infty)$ where $C(\mathcal{S}\times\mathcal{A},R)\triangleq\{F:\mathcal{S}\times\mathcal{A}\to R\}$ and $|F-G|_\infty=\sup_{s,a}|F(s,a)-G(s,a)|$*

Proof: Let $d=|F-G|_\infty$. The, $G-d\mathbf{1}\preceq F\preceq G+d\mathbf{1}$. From Lemma 3.2, $\mathcal{T}_q(G+d\mathbf{1})\succeq\mathcal{T}_qF\succeq\mathcal{T}_q(G-d\mathbf{1})$. From Lemma 3.1, $\mathcal{T}_qG+\gamma d\mathbf{1}\succeq\mathcal{T}_qF\succeq\mathcal{T}_qG-\gamma d\mathbf{1}$. Then,$\gamma d\mathbf{1}\succeq\mathcal{T}_qF-\mathcal{T}_qG\succeq-\gamma d\mathbf{1}$. Finally,

$$|\mathcal{T}_qF-\mathcal{T}_qG|_\infty\le\gamma d=\gamma|F-G|_\infty.$$

Consequently, $\mathcal{T}_q$ is γ-contraction. ■

1) Proof of Tsallis Value Iteration: *Proof of Theorem 6:* From Lemma 6.4, $\mathcal{T}_q$ is γ-contraction and has an unique fixed point $F_*=\mathcal{T}_qF_*$ from the Banach fixed point theorem. Then, for any initial function F, a sequence of F_k converges to the fixed point, i.e., $F_*=\lim_{k\to\infty}(\mathcal{T}_q)^kF_0$. The fixed point F_* satisfies a TBO equation as follows:

$$\begin{aligned} F_*(s,a) &= \mathop{\mathbb{E}}_{s'\sim P}[\mathbf{r}(s,a,s')+\gamma V_{F_*}(s')|s,a] \\ V_{F_*}(s) &= \mathop{q\text{-}\max}_{a}[F_*(s,a)], \end{aligned} \tag{50}$$

Since TBO equation is the necessary and sufficient conditions,$F_*=Q_q^\star$. ■

G. Performance Error Bounds

Theorem 7. *Let $J(\pi)$ be the expected sum of rewards of a given policy π, $\pi^\star$ be the optimal policy of an original MDP, and $\pi_q^\star$ be the optimal policy of a Tsallis MDP with an entropic index q. Then, the following inequality holds,*

$$J(\pi^\star) + (1-\gamma)^{-1} \ln_q (1/|\mathcal{A}|) \leq J(\pi_q^\star) \leq J(\pi^\star). \tag{51}$$

where $|\mathcal{A}|$ is the cardinality.

Lemma 7.1. *Let*

$$[\mathcal{T}F](s,a) \triangleq \mathop{\mathbb{E}}_{s' \sim P}[\mathbf{r}(s,a,s') + \gamma \max_{a'} F(s',a')|s,a]$$

for a function F. $\mathcal{T}$ is the original Bellman optimality operator which is used for an original value iteration. Then, for all positive integer k and any function F over $\mathcal{S} \times \mathcal{A}$,

$$\mathcal{T}_q^k F \succeq \mathcal{T}^k F$$

where $\mathcal{T}^k$ indicates k tiems application of $\mathcal{T}$. Furthermore, $V_q^\star \succeq V^\star$ holds which means that the optimal value of Tsallis MDP is greater than the optimal value of the original MDP.

Proof: When $k=1$, from Lemma 6.1, for all s, a,

$$\begin{aligned} [\mathcal{T}F](s,a) &= \mathop{\mathbb{E}}_{s' \sim P}[\mathbf{r}(s,a,s') + \gamma \max_{a'} F(s',a')|s,a] \\ &\leq \mathop{\mathbb{E}}_{s' \sim P}\left[\mathbf{r}(s,a,s') + \gamma\, q\text{-}\max_{a'} F(s',a')|s,a\right] = [\mathcal{T}_q F](s,a) \end{aligned} \tag{52}$$

Now, assume that the statement holds when $k=n$, then,

$$\mathcal{T}^{n+1}F = \mathcal{T}\mathcal{T}^n F \preceq \mathcal{T}_q\mathcal{T}^n F \preceq \mathcal{T}_q\mathcal{T}_q^n F \preceq \mathcal{T}_q^{n+1} F \tag{53}$$

From mathematical induction, the statement holds for all positive integers. Furthermore,

$$V^\star = \lim_{k\to\infty} \mathcal{T}^k F \preceq \lim_{k\to\infty} \mathcal{T}_q^k F = V_q^\star$$

■

We would like to note that the gap between $V^\star$ and $V_q^\star$ is induced from the Tsallis entropy.

1) Proof of Performance Error Bounds: *Proof of Theorem 7:* The upper bound is trivial. Since the original MDP maximizes $J(\pi)$ without the entropy maximization, it is clear that $J(\pi_q^\star) \leq J(\pi^\star)$ where $J(\pi) \triangleq \mathbb{E}_{\tau\sim\pi,P}[\sum_{t=0}^\infty \gamma^t \mathbf{R}_t]$. For the lower bound, using Lemma 7.1,

$$\begin{aligned} J(\pi^\star) = \mathbb{E}_{s_0\sim d}\left[V^\star(s_0)\right] &\leq \mathbb{E}_{s_0\sim d}\left[V_q^\star(s_0)\right] = J(\pi_q^\star) + S_q^\infty\left(\pi_q^\star\right) \\ &\leq J(\pi_q^\star) + \mathbb{E}_{\tau\sim\pi,P}\left[\sum_{t=0}^\infty \gamma^t S_q\left(\pi_q^\star(\cdot|s_t)\right)\right] \\ &\leq J(\pi_q^\star) + \mathbb{E}_{\tau\sim\pi,P}\left[\sum_{t=0}^\infty \gamma^t \max_{\pi(\cdot|s_t)} S_q\left(\pi(\cdot|s_t)\right)\right] \\ &\leq J(\pi_q^\star) - \mathbb{E}_{\tau\sim\pi,P}\left[\sum_{t=0}^\infty \gamma^t \ln_q\left(1/|\mathcal{A}|\right)\right] \\ &\leq J(\pi_q^\star) - (1-\gamma)^{-1}\ln_q\left(1/|\mathcal{A}|\right) \end{aligned} \tag{54}$$

■

H. q-Scheduling

Theorem 8 (Scheduled TPI)**.** *Let $\mathcal{TPI}_q$ be the Tsallis policy iteration operator with an entropic index q. Assume that a diverging sequence q_k is given, i.e., $\lim_{k\to\infty} q_k = \infty$. For given q_k, scheduled TPI is defined as $\mathcal{TPI}_{q_k}$, i.e., $\pi_{k+1} = \mathcal{TPI}_{q_k}(\pi_k)$. Then, $\pi_k \to \pi^\star$ as $k \to \infty$.*

Proof: The proof directly follows from Theorem 7.

$$\begin{aligned} J(\pi^\star) + (1-\gamma)^{-1}\ln_{q_k}\left(1/|\mathcal{A}|\right) &\leq J(\pi_k) \\ J(\pi^\star) + (1-\gamma)^{-1}\lim_{k\to\infty}\ln_{q_k}\left(1/|\mathcal{A}|\right) &\leq \lim_{k\to\infty} J(\pi_k) \\ J(\pi^\star) &\leq \lim_{k\to\infty} J(\pi_k) \;\; (\because \lim_{k\to\infty}\ln_{q_k}\left(1/|\mathcal{A}|\right) = 0) \\ \therefore J(\pi^\star) &= \lim_{k\to\infty} J(\pi_k) \end{aligned} \tag{55}$$

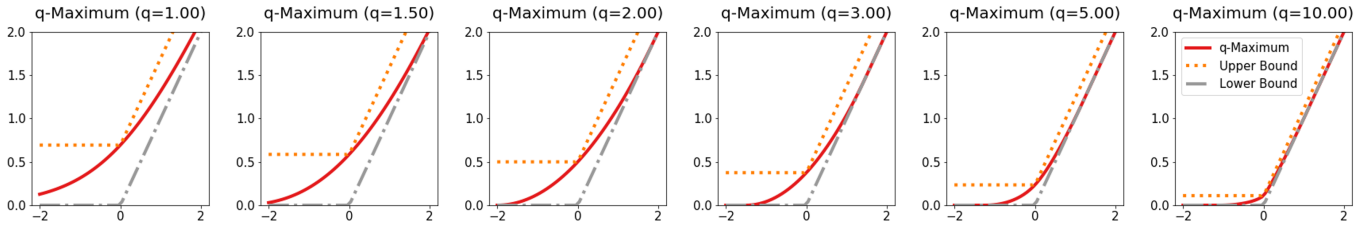


Fig. 1. Example of q-maximum operator with different q values. The figures show q-max$([c, 0])$ over $c \in [-2, 2]$. The bounds are computed by Theorem 1

■

I. Example of Bounds for q-Maximum

From theorem 1, we have the bounds for q-maximum as follows,

$$\max_x(f(x)) \leq q\text{-}\max_x(f(x)) \leq \max_x(f(x)) - \ln_q(1/|\mathcal{X}|)$$

In example, we set $\mathcal{X} = \{0, 1\}$ and $f(x)$ is deinfed as $f(0) = 0, f(1) = c$. We see the tendecy of q-maximum when c varies from -2 to 2. Then, $\max_x(f(x))$ becomes $\max([c, 0])$ and we compute the q-max$([c, 0])$ using numerical solver. Since $\mathcal{X}$ has two elements, the upper bound is $\max([c, 0]) - \ln_q(1/2)$.

Examples of q-maximum with different q values are shown in Figure 1. It can be observed that, as q increases, the bounds become tighter. Note that the gap becomes largest when $q = 1$. This gap sometimes leads to overestimation error when we use q-maximum to compute the target value of value networks.

J. Full Experimental Results on MuJoCo

The entire results are shown in Figure 2, 3, and 4. In Figure 2, average training returns of TAC with linear scheduling with different $\alpha = \{0.5, 0.2, 0.02, 0.002\}$ are shown. Note that for various α values, the optimal q is consistent. In Figure 4, average returns of every compared existing methods including PPO, TRPO, and DDPG are shown. Note that the proposed method TAC and TAC^2 consistently outperform the other actor-critic methods. We also conduct the effect of the network capacity on on-policy methods: PPO and TRPO, for fair comparison, as shown in Figure 5. We can realize that the large network capacity does not help the performance of PPO and TRPO. Therefore. this result justifies the experiments on PPO and TRPO with smaller networks in our comparison.

K. Reparameterization Trick

We follows the reparameterization trick used in the soft actor-critic method [1]. For continuous random variable, the policy network often model the Gaussian distribution where the output of π_ϕ is the mean $\mu_\phi(s)$ and standard deviation $\sigma_\phi(s)$ of Gaussian distribution. However, in most continuous control problems, the action space is often bounded. In this regards, we apply a tangent hyperbolic (tanh) to the Gaussian samples

$$a = f_\phi(s, \epsilon) = \tanh(z)$$

where

$$z = \mu_\phi(s) + \epsilon\sigma_\phi(s)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Then, the likelihood of actions $\pi_\phi(a|s)$ is computed as

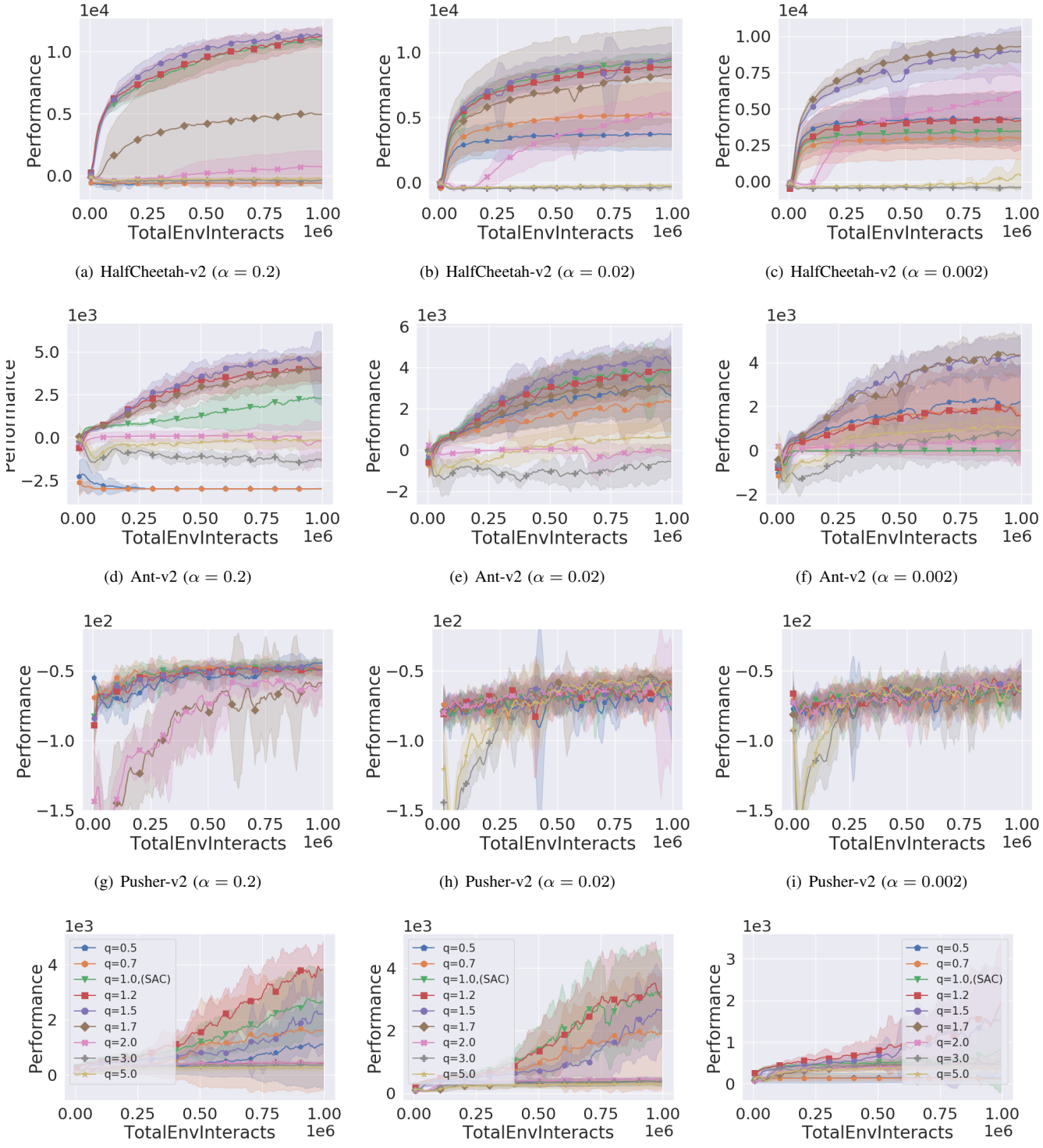
$$\pi(a|s) = \mathcal{N}\left(z; \mu_\phi(s), \sigma_\phi(s)\right) \left|\frac{\mathrm{d}a}{\mathrm{d}z}\right|^{-1}$$

where

$$\left|\frac{\mathrm{d}a}{\mathrm{d}z}\right|^{-1} = \Pi_{i=1}^{D}(1 - \tanh(z_i))^{-1}$$

where D is a dimension of z or a and z_i is the ith element of z. Finally, the q-logarithm of the policy distribution is

$$\ln_q\left(\mathcal{N}\left(z; \mu_\phi(s), \sigma_\phi(s)\right) \left|\frac{\mathrm{d}a}{\mathrm{d}z}\right|^{-1}\right)$$



(a) HalfCheetah-v2 ($\alpha = 0.2$)

(b) HalfCheetah-v2 ($\alpha = 0.02$)

(c) HalfCheetah-v2 ($\alpha = 0.002$)

(d) Ant-v2 ($\alpha = 0.2$)

(e) Ant-v2 ($\alpha = 0.02$)

(f) Ant-v2 ($\alpha = 0.002$)

(g) Pusher-v2 ($\alpha = 0.2$)

(h) Pusher-v2 ($\alpha = 0.02$)

(i) Pusher-v2 ($\alpha = 0.002$)

(j) Humanoid-v2 ($\alpha = 0.05$)

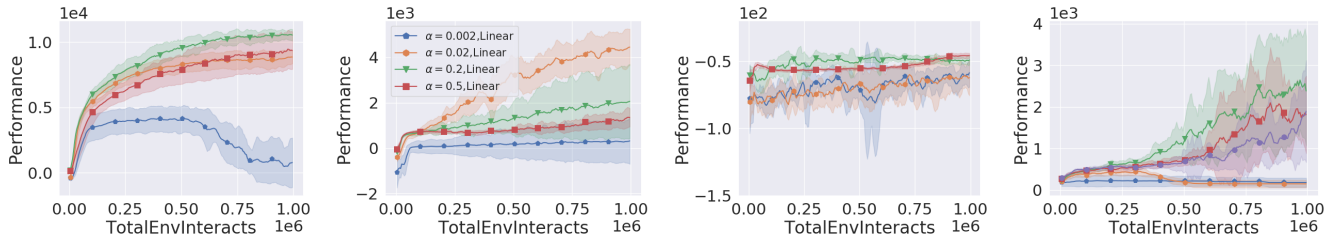
(k) Humanoid-v2 ($\alpha = 0.02$)

(l) Humanoid-v2 ($\alpha = 0.002$)

Fig. 2. Average training returns on MuJoCo environments. Tsallis actor-critics with q values : $1.0, 1.2, 1, 5$ and 1.7 and $\alpha = \{0.2, 0.02, 0.002\}$ generally show better performance than other entropic indices. Real line is an average return over ten trials and shade area shows a variance.

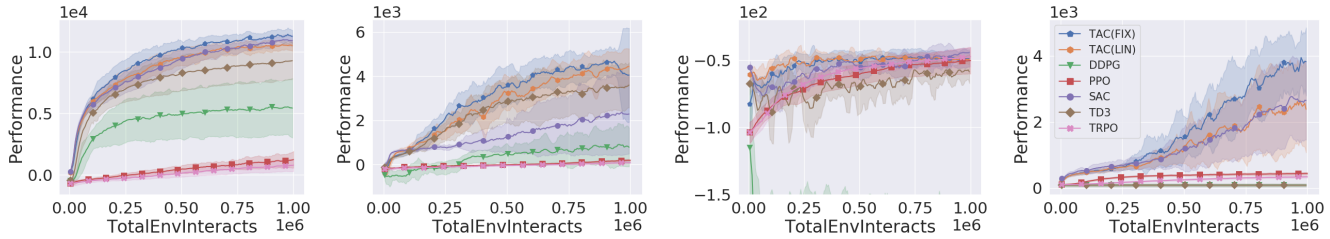
L. Numerical Issue

Since we handle the continuous action space, the policy is modeled as a density function of a continuous random variable. Unlikely to a probability mass function, a probability density function (pdf) sometimes diverges to infinity (or the maximum number of a computing machine) when the probability is concentrated at a single point. In this case, the large pdf value induces a large gradient which makes the learning procedure unstable. Thus, in our implementation, the pdf value is restricted to no



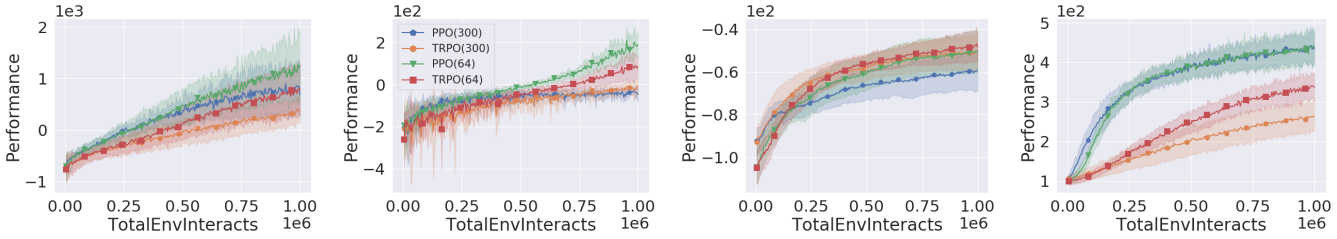
(a) HalfCheetah-v2 (b) Ant-v2 (c) Pusher-v2 (d) Humanoid-v2

Fig. 3. Average training returns of TAC with linear scheduling with different $\alpha = \{0.5, 0.2, 0.02, 0.002\}$.



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

Fig. 4. Comparison to existing actor-critic methods on four MuJoCo environments. Soft actor-critic (red square line) is the same as Tsallis actorc-critic with $q = 1$ and TAC (blue pentagon line) indicates Tasllis actor-critic with $q \neq 1$.



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

Fig. 5. Average training returns of PPO and TRPO on MuJoCo environments. The number in parentheses indicates the number of hidden units. Both algorithms have better performance when using smaller networks

greater than $10^{\frac{8}{q-1}}$. This numerical issue is often caused when $q \geq 2$. For $q \geq 2$, we use the following likelihood.

$$\ln_q \left(\min \left(10^{\frac{8}{q-1}}, \mathcal{N}\left(z; \mu_\phi(s), \sigma_\phi(s)\right) \left| \frac{\mathrm{d}a}{\mathrm{d}z} \right|^{-1} \right) \right) \quad \left(\leq \frac{10^8 - 1}{q - 1} \right)$$

M. Hyperparameter Settings

TABLE I
HYPERPARAMETER OF TSALLIS ACTOR CRITIC

Parameter	Value
Optimizer	Adam in TensorFlow
Learning rate	$1e^{-3}$
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	$400, 300$ (MuJoCo) and 100 (Soft Mobile Robot)
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

TABLE II
HYPERPARAMETER FOR FIXED q

Environment	(Best) Entropy Coefficient, α	(Best) Entropic Index, q
HalfCheetah-v2	0.2	1.5
Ant-v2	0.2	1.5
Pusher-v2	0.2	1.2 (slightly better)
Humanoid-v2	0.05	1.2

TABLE III
HYPERPARAMETER FOR LINEAR SCHEDULING

Environment	(Best) Entropy Coefficient, α
HalfCheetah-v2	0.2
Ant-v2	0.2
Pusher-v2	0.02
Humanoid-v2	0.02

REFERENCES

[1] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 1856–1865, Stockholmsmässan, Stockholm, Sweden, 2018.

[2] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[3] Umar Syed, Michael H. Bowling, and Robert E. Schapire. Apprenticeship learning using linear programming. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008)*, pages 1032–1039, Helsinki, Finland, 2008.