A Basic facts of Tsallis KL divergence

We present some basic facts about q-logarithm and Tsallis KL divergence.

We begin by introducing the $2-q$ duality for Tsallis statistics. Recall that the q-logarithm and Tsallis entropy defined in the main paper are:

$$\ln_q x = \frac{x^{1-q}-1}{1-q}, \quad S_q(x) = -\left\langle x^q, \ln_q x \right\rangle .$$

In the RL literature, another definition $q^* = 2-q$ is more often used [Lee et al., 2020]. This is called the $2-q$ duality [Naudts, 2002, Suyari and Tsukada, 2005], which refers to that the Tsallis entropy can be equivalently defined as:

$$\ln_{q^*} x = \frac{x^{q^*-1}-1}{q^*-1}, \quad S_{q^*}(x) = -\left\langle x, \ln_{q^*} x \right\rangle ,$$

By the duality we can show [Suyari and Tsukada, 2005, Eq.(12)]:

$$S_q(x) := -\left\langle x^q, \frac{x^{1-q}-1}{1-q} \right\rangle = \frac{\langle \mathbf{1}, x^q\rangle - 1}{1-q} = \frac{\left\langle \mathbf{1}, x^{q^*}\right\rangle - 1}{1-q^*} = -\left\langle x, \frac{x^{q^*-1}-1}{q^*-1} \right\rangle =: S_{q^*}(x),$$

i.e. the duality between logarithms $\ln_{q^*} x$ and $\ln_q x$ allows us to define Tsallis entropy by an alternative notation q^* that eventually reaches to the same functional form.

We now come to examine Tsallis KL divergence (or Tsallis relative entropy) defined in another form: $D^q_{KL}(\pi \,\|\, \mu) = \left\langle \pi, \ln_{q^*} \frac{\pi}{\mu} \right\rangle$ [Prehl et al., 2012]. In the main paper we used the definition $D^q_{KL}(\pi \,\|\, \mu) = \left\langle \pi, -\ln_q \frac{\mu}{\pi} \right\rangle$ [Furuichi et al., 2004]. We show they are equivalent by the same logic:

$$\left\langle \pi, -\ln_q \frac{\mu}{\pi} \right\rangle = \left\langle \pi, -\frac{\left(\frac{\mu}{\pi}\right)^{1-q} - 1}{1-q} \right\rangle = \left\langle \pi, \frac{\left(\frac{\pi}{\mu}\right)^{q-1} - 1}{q-1} \right\rangle = \left\langle \pi, \ln_{q^*} \frac{\pi}{\mu} \right\rangle . \tag{12}$$

The equivalence allows us to work with whichever of $\ln_q$ and $\ln_{q^*}$ that makes the proof easier to work out the following useful properties of Tsallis KL divergence:

— Nonnegativity $D^q_{KL}(\pi \,\|\, \mu) \geq 0$: since the function $-\ln_q \pi$ is convex, by Jensen's inequality

$$\left\langle \pi, -\ln_q \frac{\mu}{\pi} \right\rangle \geq -\ln_q \left\langle \pi, \frac{\mu}{\pi} \right\rangle = 0,$$

— Conditions of $D^q_{KL}(\pi \,\|\, \mu) = 0$: directly from the above, in Jensen's inequality the equality holds only when $\frac{\mu}{\pi} = 1$ almost everywhere, i.e. $D^q_{KL}(\pi \,\|\, \mu) = 0$ implies $\mu = \pi$ almost everywhere.

— Conditions of $D^q_{KL}(\pi \,\|\, \mu) = \infty$: To better align with the standard KL divergence, let us work with $\ln_{q^*}$, following [Cover and Thomas, 2006], let us define

$$0 \ln_{q^*} \frac{0}{0} = 0, \quad 0 \ln_{q^*} \frac{0}{\mu} = 0, \quad \pi \ln_{q^*} \frac{\pi}{0} = \infty.$$

We conclude that $D^q_{KL}(\pi \,\|\, \mu) = \infty$ whenever $\pi > 0$ and $\mu = 0$.

— Bounded entropy $\forall q$, $0 \leq S_q(\pi) \leq \ln_q |\mathcal{A}|$: let $\mu = \frac{1}{|\mathcal{A}|}$, by the nonnegativity of Tsallis KL divergence:

$$D^q_{KL}(\pi \,\|\, \mu) = \left\langle \pi, -\ln_q \frac{1}{(|\mathcal{A}| \cdot \pi)} \right\rangle = \left\langle \pi, \frac{(|\mathcal{A}| \cdot \pi)^{q-1} - 1}{q-1} \right\rangle = |\mathcal{A}|^{q-1} \left(\frac{\langle \mathbf{1}, \pi^q \rangle - 1}{q-1} - \frac{\frac{1}{|\mathcal{A}|^{q-1}} - 1}{q-1} \right) \geq 0.$$

Notice that $\frac{\langle \mathbf{1}, \pi^q\rangle - 1}{q-1} = \left\langle \pi^q, \frac{1-\pi^{1-q}}{1-q} \right\rangle = \langle \pi, \ln_q \pi \rangle = -S_q(\pi)$ and $\frac{\frac{1}{|\mathcal{A}|^{q-1}} - 1}{q-1} = \ln_q |\mathcal{A}|$, we conclude

$$S_q(\pi) \leq \ln_q |\mathcal{A}|.$$

B Proof of Theorem 1 and 2

We structure this section as the following three parts:

1. Tsallis entropy regularized policy has general expression for all q. Moreover, q and τ are interchangeable for controlling the truncation (Theorem 1).
2. The policies can be expressed by q-exponential (Theorem 1).
3. We present a computable approximate threshold $\hat{\psi}_q$ (Theorem 2).

General expression for Tsallis entropy regularized policy. The original definition of Tsallis entropy is $S_{q^*}(\pi(\cdot|s)) = \frac{p}{q^*-1}\left(1 - \sum_a \pi^{q^*}(a|s)\right), q^* \in \mathbb{R},\ p \in \mathbb{R}_+$. Note that similar to Appendix A, we can choose whichever convenient of q and q^*, since the domain of the entropic index is $\mathbb{R}$. To obtain the Tsallis entropy-regularized policies we follow [Chen et al., 2018]. The derivation begins with assuming an actor-critic framework where the policy network is parametrized by w. It is well-known that the parameters should be updated towards the direction specified by the policy gradient theorem:

$$\Delta w \propto \mathbb{E}_\pi\left[Q_\pi \frac{\partial \ln \pi}{\partial w} + \tau \frac{\partial \mathcal{H}(\pi)}{\partial w}\right] - \sum_s \lambda(s) \frac{\partial \langle \mathbf{1}, \pi\rangle}{\partial w} =: f(w), \tag{13}$$

Recall that $\mathcal{H}(\pi)$ denotes the Shannon entropy and τ is the coefficient. $\lambda(s)$ are the Lagrange multipliers for the constraint $\langle \mathbf{1}, \pi\rangle = 1$. In the Tsallis entropy framework, we replace $\mathcal{H}(\pi)$ with $S_{q^*}(\pi)$. We can assume $p = \frac{1}{q^*}$ to ease derivation, which is the case for sparsemax.

We can now explicitly write the optimal condition for the policy network parameters:

$$\begin{aligned} f(w) = 0 &= \mathbb{E}_\pi\left[Q_\pi \frac{\partial \ln \pi}{\partial w} + \tau \frac{\partial S_{q^*}(\pi)}{\partial w}\right] - \sum_s \lambda(s) \frac{\partial \langle \mathbf{1}, \pi\rangle}{\partial w} \\ &= \mathbb{E}_\pi\left[Q_\pi \frac{\partial \ln \pi}{\partial w} - \tau \frac{1}{q^*-1}\left\langle \mathbf{1}, \pi^{q^*} \frac{\partial \ln \pi}{\partial w}\right\rangle - \tilde{\psi}_q(s) \frac{\partial \ln \pi}{\partial w}\right] \\ &= \mathbb{E}_\pi\left[\left(Q_\pi - \tau \frac{1}{q^*-1}\pi^{q^*-1} - \tilde{\psi}_q(s)\right)\frac{\partial \ln \pi}{\partial w}\right], \end{aligned} \tag{14}$$

where we leveraged $\frac{\partial S_{q^*}(\pi)}{\partial w} = \frac{1}{q^*-1}\left\langle \mathbf{1}, \pi^{q^*} \frac{\partial \ln \pi}{\partial w}\right\rangle$ in the second step and absorbed terms into the expectation in the last step. $\tilde{\psi}_q(s)$ denotes the adjusted Lagrange multipliers by taking $\lambda(s)$ inside the expectation and modifying it according to the discounted stationary distribution.

Now it suffices to verify either $\frac{\partial \ln \pi}{\partial w} = 0$ or

$$\begin{aligned} & Q_\pi(s,a) - \tau \frac{1}{q^*-1}\pi^{q^*-1}(a|s) - \tilde{\psi}_q(s) = 0 \\ \Leftrightarrow \quad & \pi^*(a|s) = \sqrt[q^*-1]{\left[\frac{Q_\pi(s,a)}{\tau} - \frac{\tilde{\psi}_q(s)}{\tau}\right]_+ (q^*-1)}, \\ \text{or} \quad & \pi^*(a|s) = \sqrt[1-q]{\left[\frac{Q_\pi(s,a)}{\tau} - \frac{\tilde{\psi}_q(s)}{\tau}\right]_+ (1-q)}, \end{aligned} \tag{15}$$

where we changed the entropic index from q^* to q. Clearly, the root does not affect truncation. Consider the pair $(q^* = 50, \tau)$, then the same truncation effect can be achieved by choosing $(q^* = 2, \frac{\tau}{50-1})$. The same goes for q. Therefore, we conclude that q and τ are interchangeable for the truncation, and we should stick to the analytic choice $q^* = 2 (q = 0)$.

Tsallis policies can be expressed by q-exponential. Given Eq. (15), by adding and subtracting 1, we have:

$$\pi^*(a|s) = \sqrt[1-q]{\left[1 + (1-q)\left(\frac{Q_\pi(s,a)}{\tau} - \tilde{\psi}_q\left(\frac{Q_\pi(s,\cdot)}{\tau}\right) - \frac{1}{1-q}\right)\right]_+} = \exp_q\left(\frac{Q_\pi(s,a)}{\tau} - \hat{\psi}_q\left(\frac{Q_\pi(s,\cdot)}{\tau}\right)\right),$$

where we defined $\hat{\psi}_q = \tilde{\psi}_q + \frac{1}{1-q}$. Note that this expression is general for all q, but whether π^* has closed-form expression depends on the solvability of $\tilde{\psi}_q$.

Let us consider the extreme case $q = \infty$. It is clear that $\lim_{q\to\infty} \frac{1}{1-q} \to 0$. Therefore, for any $x > 0$ we must have $x^{\frac{1}{1-q}} \to 1$; i.e., there is only one action with probability 1, with all others being 0. This conclusion agrees with the fact that $S_q(\pi) \to 0$ as $q \to \infty$: hence the regularized policy degenerates to $\arg\max$.

A computable Normalization Function. The constraint $\sum_{a\in K(s)} \pi^*(a|s) = 1$ is exploited to obtain the threshold ψ for the sparsemax [Lee et al., 2018, Chow et al., 2018]. Unfortunately, this is only possible when the root vanishes, since otherwise the constraint yields a summation of radicals. Nonetheless, we can resort to first-order Taylor expansion for deriving an approximate policy. Following [Chen et al., 2018], let us expand Eq. (15) by the first order Taylor expansion $f(z) + f'(z)(x - z)$, where we let $z = 1$, $x = \left[\frac{Q_\pi(s,a)}{\tau} - \tilde{\psi}_q\left(\frac{Q_\pi(s,\cdot)}{\tau}\right)\right]_+ (1-q)$, $f(x) = x^{\frac{1}{1-q}}$, $f'(x) = \frac{1}{1-q}x^{\frac{q}{1-q}}$. So that the unnormalized approximate policy has

$$\begin{aligned}\tilde{\pi}^*(a|s) &\approx f(z) + f'(z)(x - z) \\ &= 1 + \frac{1}{1-q}\left(\left(\frac{Q_\pi(s,a)}{\tau} - \tilde{\psi}_q\left(\frac{Q_\pi(s,\cdot)}{\tau}\right)\right)(1-q) - 1\right).\end{aligned} \tag{16}$$

Therefore it is clear as $q \to \infty, \tilde{\pi}^*(a|s) \to 1$. This concords well with the limit case where $\pi^*(a|s)$ degenerates to $\arg\max$. With Eq. (16), we can solve for the approximate normalization by the constraint $\sum_{a\in K(s)} \pi^*(a|s) = 1$:

$$\begin{aligned}1 &= \sum_{a\in K(s)} \left[1 + \frac{1}{1-q}\left(\left(\frac{Q_\pi(s,a)}{\tau} - \tilde{\psi}_q\left(\frac{Q_\pi(s,\cdot)}{\tau}\right)\right)(1-q) - 1\right)\right] \\ &= |K(s)| - \frac{1}{1-q}|K(s)| + \sum_{a\in K(s)}\left[\frac{Q_\pi(s,a)}{\tau} - \tilde{\psi}_q\left(\frac{Q_\pi(s,\cdot)}{\tau}\right)\right] \\ \Leftrightarrow \tilde{\psi}_q\left(\frac{Q_\pi(s,\cdot)}{\tau}\right) &= \frac{\sum_{a\in K(s)}\frac{Q_\pi(s,\cdot)}{\tau} - 1}{|K(s)|} + 1 - \frac{1}{1-q}.\end{aligned}$$

In order for an action to be in $K(s)$, it has to satisfy $\frac{Q_\pi(s,\cdot)}{\tau} > \frac{\sum_{a\in K(s)}\frac{Q_\pi(s,\cdot)}{\tau} - 1}{|K(s)|} + 1 - \frac{1}{1-q}$. Therefore, the condition of $K(s)$ satisfies:

$$1 + i\frac{Q_\pi(s,a_{(i)})}{\tau} > \sum_{j=1}^{i}\frac{Q_\pi(s,a_{(j)})}{\tau} + i\left(1 - \frac{1}{1-q}\right).$$

Therefore, we see the approximate threshold $\hat{\psi}_q = \tilde{\psi}_q + 1$. When $q = 0$ or $q^* = 2$, $\hat{\psi}_q$ recovers ψ and hence $\tilde{\pi}^*$ recovers the exact sparsemax policy.

C Proof of convergence of $\Omega(\pi) = D^q_{KL}(\pi \,\|\, \cdot)$ when $q = 2$

Let us work with $\ln_{q^*}$ from Appendix A and define $\|\cdot\|_p$ as the l_p-norm. The convergence proof for $\Omega(\pi) = D^q_{KL}(\pi \,\|\, \cdot)$ when $q = 2$ comes from that $\Omega(\pi)$ is strongly convex in π:

$$\Omega(\pi) = D_{KL}^{q^*=2}(\pi||\cdot) = \left\langle \pi, \ln_2\frac{\pi}{\cdot}\right\rangle = \left\langle \pi, \frac{\left(\frac{\pi}{\cdot}\right)^{2-1} - 1}{2-1}\right\rangle \propto \left\|\frac{\pi}{\cdot}\right\|_2^2 - 1. \tag{17}$$

Similarly, the negative Tsallis sparse entropy $-S_2(\pi)$ is also strongly convex. Then the propositions of [Geist et al., 2019] can be applied, which we restate in the following:

Lemma 1 ([Geist et al., 2019]). *Define regularized value functions as:*

$$Q_{\pi,\Omega} = r + \gamma P V_{\pi,\Omega}, \qquad V_{\pi,\Omega} = \langle \pi, Q_{\pi,\Omega}\rangle - \Omega(\pi).$$

If $\Omega(\pi)$ is strongly convex, let $\Omega^*(Q) = \max_\pi \langle \pi, Q\rangle - \Omega(\pi)$ denote the Legendre-Fenchel transform of $\Omega(\pi)$, then

Algorithm 1: MVI(q)

```
Input: number of iterations $T$, entropy coefficient $\tau$, TKL coefficient $\alpha$
Initialize $Q_0, \pi_0$ arbitrarily
Let $\{|\mathcal{A}|\} = \{1, 2, \dots, |\mathcal{A}|\}$
for $k = 1, 2, \dots, T$ do
  # Policy Improvement
  for $(s, a) \in (\mathcal{S}, \mathcal{A})$ do
    Sort $Q_k(s, a_{(1)}) > \cdots > Q_k(s, a_{(|\mathcal{A}|)})$
    Find $K(s) = \max\left\{ i \in \{|\mathcal{A}|\} \,\middle|\, 1 + i\frac{Q_k(s,a_{(i)})}{\tau} > \sum_{j=1}^{i} \frac{Q_k(s,a_{(j)})}{\tau} + i\left(1 - \frac{1}{1-q}\right) \right\}$
    Compute $\hat{\psi}_q\left(\frac{Q_k(s,\cdot)}{\tau}\right) = \frac{\sum_{a\in K(s)} \frac{Q_k(s,a)}{\tau} - 1}{|K(s)|} + 1$
    # Normalize when $q \neq 2$
    $\pi_{k+1}(a|s) \propto \exp_q\left(\frac{Q_k(s,a)}{\tau} - \hat{\psi}_q\left(\frac{Q_k(s,\cdot)}{\tau}\right)\right)$
  end for
  # Policy Evaluation
  for $(s, a, s') \in (\mathcal{S}, \mathcal{A})$ do
    $Q_{k+1}(s, a) =$
    $r(s,a) + \alpha\tau\left(Q_k(s,a) - \mathcal{M}_{q,\tau} Q_k(s)\right) + \gamma \sum_{b\in\mathcal{A}} \pi_{k+1}(b|s')\left(Q_k(s', b) - \tau \ln_q \pi_{k+1}(b|s')\right)$
  end for
end for
```

- $\nabla\Omega^*$ *is Lipschitz and is the unique maximizer of* $\arg\max_\pi \langle \pi, Q\rangle - \Omega(\pi)$.
- $T_{\pi,\Omega}$ *is a* γ*-contraction in the supremum norm, i.e.* $\|T_{\pi,\Omega}V_1 - T_{\pi,\Omega}V_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$. *Further, it has a unique fixed point* $V_{\pi,\Omega}$.
- *The policy* $\pi_{*,\Omega} = \arg\max_\pi \langle \pi, Q_{*,\Omega}\rangle - \Omega(\pi)$ *is the unique optimal regularized policy.*

Note that in the main paper we dropped the subscript Ω for both the regularized optimal policy and action value function to lighten notations. It is now clear that Eq. (6) indeed converges for entropic indices that make $D^q_{KL}(\pi \,\|\, \cdot)$ strongly convex. But we mostly consider the case $q = 2$.

D Derivation of the Tsallis KL Policy

D.1 Tsallis KL Policies are Similar to KL

We extend the proof and use the same notations from [Lee et al., 2020, Appendix D] to derive the Tsallis KL regularized policy. Again let us work with $\ln_{q^*}$ from A. Define state visitation as $\rho_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \mathbb{1}(s_t = s)\right]$ and state-action visitaion $\rho_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \mathbb{1}(s_t = s, a_t = a)\right]$. The core of the proof resides in establishing the one-to-one correspondence between the policy and the induced state-action visitation ρ_π. For example, Tsallis entropy is written as

$$S_{q^*}(\pi) = S_{q^*}(\rho_\pi) = -\sum_{s,a} \rho_\pi(s,a) \ln_{q^*} \frac{\rho_\pi(s,a)}{\sum_a \rho_\pi(s,a)}.$$

This unique correspondence allows us to replace the optimization variable from π to ρ_π. Indeed, one can always restore the policy by $\pi(a|s) := \frac{\rho_\pi(s,a)}{\sum_{a'} \rho_\pi(s,a')}$.

Let us write Tsallis KL divergence as $D^{q^*}_{KL}(\pi \,\|\, \mu) = D^{q^*}_{KL}(\rho \,\|\, \nu) = \sum_{s,a} \rho(s,a) \ln_{q^*} \frac{\rho(s,a) \sum_{a'} \nu(s,a')}{\nu(s,a) \sum_{a'} \rho(s,a')}$ by replacing the policies π, μ with their state-action visitation ρ, ν. One can then convert the Tsallis MDP problem into the following problem:

$$\begin{aligned}
&\max_{\rho} \sum_{s,a} \rho(s,a) \sum_{s'} r(s,a) P(s'|s,a) - D_{KL}^{q^*}(\rho \,\|\, \nu) \\
&\text{subject to } \forall s, a, \quad \rho(s,a) > 0, \\
&\sum_{a} \rho(s,a) = d(s) + \sum_{s',a'} P(s|s',a') \rho(s',a'),
\end{aligned} \tag{18}$$

where $d(s)$ is the initial state distribution. Eq. (18) is known as the Bellman Flow Constraints [Lee et al., 2020, Prop. 5] and is concave in ρ since the first term is linear and the second term is concave in ρ. Then the primal and dual solutions satisfy KKT conditions sufficiently and necessarily. Following [Lee et al., 2020, Appendix D.2], we define the Lagrangian objective as

$$\begin{aligned}
\mathcal{L} := &\sum_{s,a} \rho(s,a) \sum_{s'} r(s,a) P(s'|s,a) - D_{KL}^{q^*}(\rho \,\|\, \nu) + \sum_{s,a} \lambda(s,a) \rho(s,a) \\
&+ \sum_{s} \zeta(s) \left(d(s) + \sum_{s',a'} P(s|s',a') \rho(s',a') - \sum_{a} \rho(s,a) \right)
\end{aligned}$$

where $\lambda(s,a)$ and $\zeta(s)$ are dual variables for nonnegativity and Bellman flow constraints. The KKT conditions are:

$$\begin{aligned}
\forall s, a, \quad \rho^*(s,a) &\geq 0, \\
d(s) + \sum_{s',a'} P(s|s',a') \rho^*(s',a') - \sum_{a} \rho^*(s,a) &= 0, \\
\lambda^*(s,a) \leq 0, \quad \lambda^*(s,a) \rho^*(s,a) &= 0, \\
0 = \sum_{s'} r(s,a) P(s'|s,a) + \gamma \sum_{s'} \zeta^*(s') P(s'|s,a) - \zeta^*(s) + \lambda^*(s,a) &- \frac{\partial D_{KL}^{q^*}(\rho^* \,\|\, \nu)}{\partial \rho(s,a)}, \\
\text{where } -\frac{\partial D_{KL}^{q^*}(\rho^* \,\|\, \nu)}{\partial \rho(s,a)} &= -\ln_{q^*} \frac{\rho^*(s,a) \sum_{a'} \nu(s,a')}{\nu(s,a) \sum_{a'} \rho^*(s,a')} - \left(\frac{\rho^*(s,a) \sum_{a'} \nu(s,a')}{\nu(s,a) \sum_{a'} \rho^*(s,a')} \right)^{q^*-1} \\
&\quad + \sum_{a} \left(\frac{\rho^*(s,a)}{\sum_{a'} \rho^*(s,a')} \right)^{q^*} \left(\frac{\sum_{a'} \nu(s,a)}{\nu(s,a)} \right)^{q^*-1}.
\end{aligned}$$

The dual variable $\zeta^*(s)$ can be shown to equal to the optimal state value function $V^*(s)$ following Lee et al. [2020], and $\lambda^*(s,a) = 0$ whenever $\rho^*(s,a) > 0$.

By noticing that $x^{q^*-1} = (q^* - 1) \ln_{q^*} x + 1$, we can show that $-\frac{\partial D_{KL}^{q^*}(\rho^* \| \nu)}{\partial \rho(s,a)} = -q^* \ln_{q^*} \frac{\rho^*(s,a) \sum_{a'} \nu(s,a')}{\nu(s,a) \sum_{a'} \rho^*(s,a')} - 1 + \sum_a \left(\frac{\rho^*(s,a)}{\sum_{a'} \rho^*(s,a')} \right)^{q^*} \left(\frac{\sum_{a'} \nu(s,a)}{\nu(s,a)} \right)^{q^*-1}$. Substituting $\zeta^*(s) = V^*(s)$, $\pi^*(a|s) = \frac{\rho^*(s,a)}{\sum_{a'} \rho^*(s,a)}$, $\mu^*(a|s) = \frac{\nu^*(s,a)}{\sum_{a'} \nu^*(s,a)}$ into the above KKT condition and leverage the equality $Q^*(s,a) = r(s,a) + \mathbb{E}_{s' \sim P}[\gamma \zeta^*(s')]$ we have:

$$\begin{aligned}
&Q^*(s,a) - V^*(s) - q^* \ln_{q^*} \frac{\pi(a|s)}{\mu(a|s)} - 1 + \sum_{a'} \pi(a|s) \left(\frac{\pi(a|s)}{\mu(a|s)} \right)^{q^*-1} = 0 \\
&\Leftrightarrow \pi^*(a|s) = \mu(a|s) \exp_{q^*} \left(\frac{Q^*(s,a)}{q^*} - \frac{V^*(s) + 1 - \sum_{a'} \pi(a|s) \left(\frac{\pi(a|s)}{\mu(a|s)} \right)^{q^*-1}}{q^*} \right).
\end{aligned}$$

By comparing it to the maximum Tsallis entropy policy [Lee et al., 2020, Eq.(49)] we see the only difference lies in the baseline term $\mu(a|s)^{-(q^*-1)}$, which is expected since we are exploiting Tsallis KL regularization. Let us define the normalization function as

$$\psi\left(\frac{Q^*(s,\cdot)}{q^*} \right) = \frac{V^*(s) + 1 - \sum_a \pi(a|s) \left(\frac{\pi(a|s)}{\mu(a|s)} \right)^{q^*-1}}{q^*},$$

then we can write the policy as

$$\pi^*(a|s) = \mu(a|s) \exp_{q^*}\left(\frac{Q^*(s,a)}{q^*} - \psi\left(\frac{Q^*(s,\cdot)}{q^*}\right)\right).$$

In a way similar to KL regularized policies, at $k+1$-th update, take $\pi^* = \pi_{k+1}, \mu = \pi_k$ and $Q^* = Q_k$, we write $\pi_{k+1} \propto \pi_k \exp_q Q_k$ since the normalization function does not depend on actions. We ignored the scaling constant q^* and regularization coefficient. Hence one can now expand Tsallis KL policies as:

$$\pi_{k+1} \propto \pi_k \exp_{q^*}(Q_k) \propto \pi_{k-1} \exp_{q^*}(Q_{k-1}) \exp_{q^*}(Q_k) \propto \cdots \propto \exp_{q^*} Q_1 \cdots \exp_{q^*} Q_k,$$

which proved the first part of Eq. (7).

D.2 Tsallis KL Policies Do More than Average

We now show the second part of Eq. (7), which stated that the Tsallis KL policies do more than average. This follows from the following lemma:

Lemma 2 (Eq. (25) of [Yamano, 2002]).

$$\left(\exp_q x_1 \ldots \exp_q x_n\right)^{1-q} = \exp_q\left(\sum_{j=1}^{k} x_j\right)^{1-q} + \sum_{j=2}^{k}(1-q)^j \sum_{i_1=1<\cdots<i_j}^{k} x_{i_1}\cdots x_{i_j}. \quad (19)$$

However, the mismatch between the base q and the exponent $1-q$ is inconvenient. We exploit the $q = 2 - q^*$ duality to show this property holds for q^* as well:

$$\begin{aligned}\left(\exp_{q^*} x \cdot \exp_{q^*} y\right)^{q^*-1} &= [1+(q^*-1)x]_+ \cdot [1+(q^*-1)y]_+ \\ &= \left[1+(q^*-1)x + (q^*-1)y + (q^*-1)^2 xy\right]_+ \\ &= \exp_q(x+y)^{q^*-1} + (q^*-1)^2 xy.\end{aligned}$$

Now since we proved the two-point property for q^*, by the same induction steps in [Yamano, 2002, Eq. (25)] we conclude the proof. The weighted average part Eq. (8) comes immediately from [Suyari et al., 2020, Eq.(18)].

E Implementation Details

We list the hyperparameters for Gym environments in Table 1. The epsilon threshold is fixed at 0.01 from the beginning of learning. FC n refers to the fully connected layer with n activation units.

For the Atari games we implemented MVI(q), Tsallis-VI and M-VI based on the Quantile Regression DQN Dabney et al. [2018]. We leverage the optimized Stable-Baselines3 architecture [Raffin et al., 2021] for best performance. The details can be seen from Table 2. The Q-network uses 3 convolutional layers. The epsilon greedy threshold is initialized at 1.0 and gradually decays to 0.01 at the end of first 10% of learning. For conservative learning, we choose the Tsallis entropy coefficient as $\alpha = 10$.

We show in Figure 6 the full learning curves of MVI(q). Figures 7 and 8 show the full learning curves of Tsallis-VI.

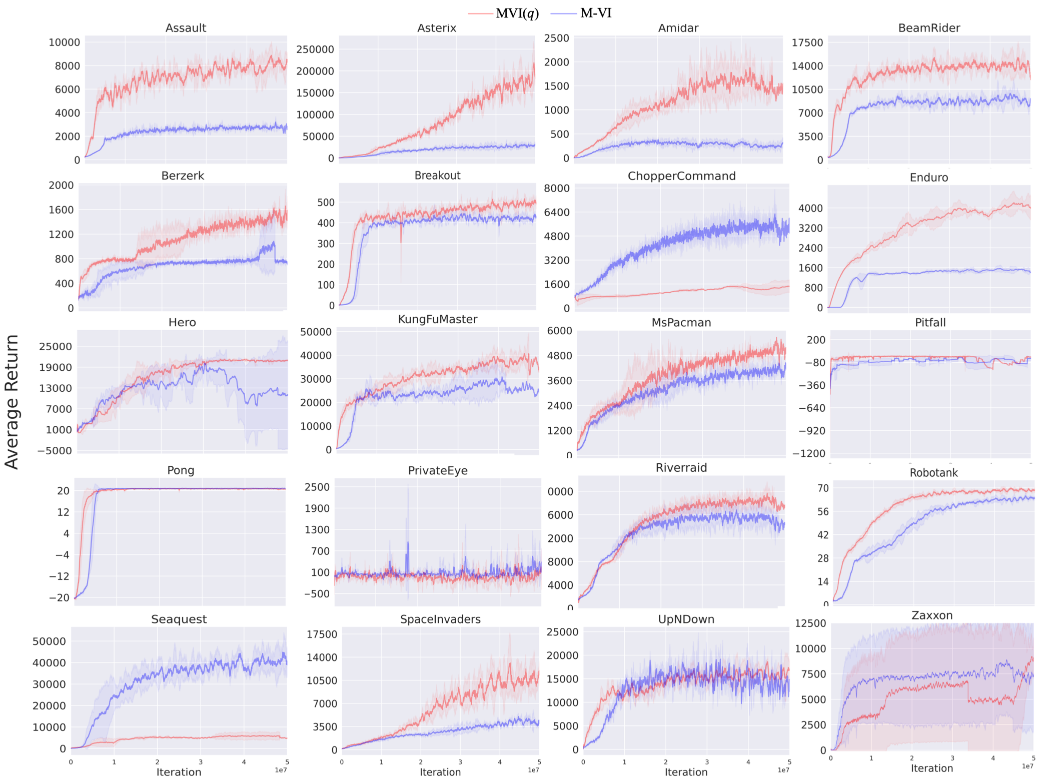

Figure 6: Learning curves of MVI(q) and M-VI on the selected Atari games.

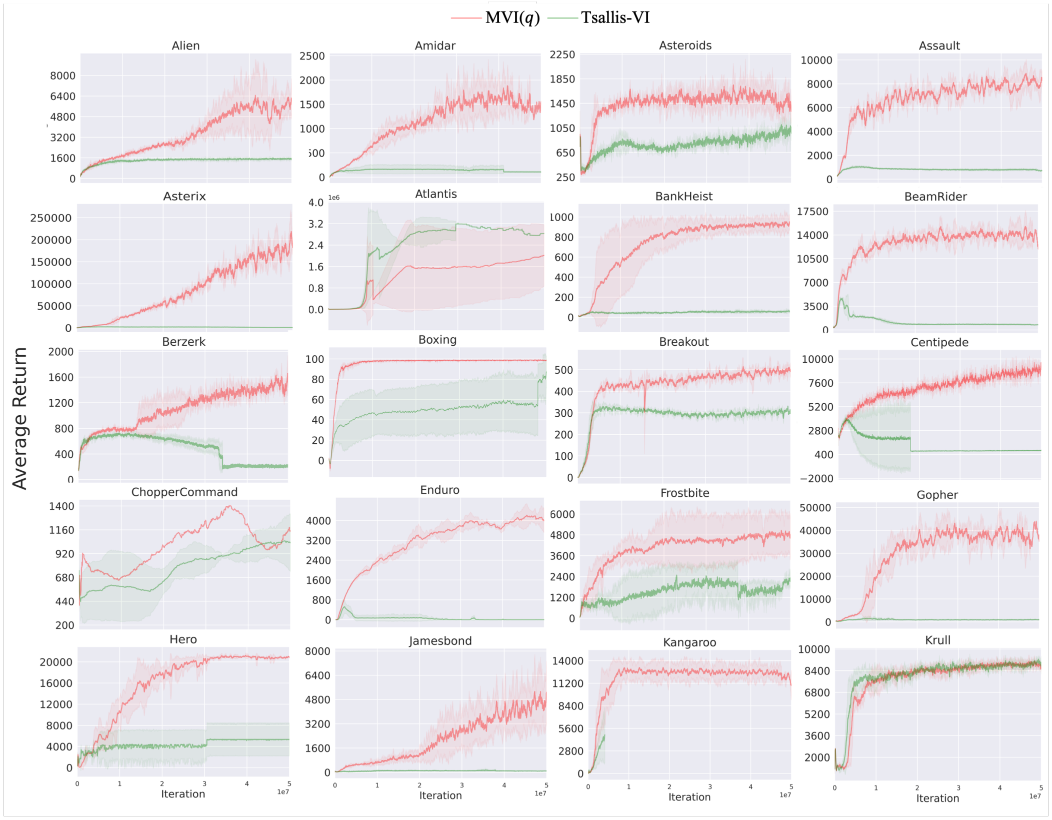

Figure 7: Learning curves of MVI(q) and Tsallis-VI on the selected Atari games.

Table 1: Parameters used for CartPole-v1.

Network Parameter	Value	Algorithm Parameter	Value
T (total steps)	5×10^5	γ (discount rate)	0.99
C (interaction period)	4	ϵ (epsilon greedy threshold)	0.01
$\lvert B \rvert$ (buffer size)	5×10^4	τ (Tsallis entropy coefficient)	0.03
B_t (batch size)	128	α (advantage coefficient)	0.9
I (update period)	100 (Car.) / 2500 (Others)		
Q-network architecture	FC512 - FC512		
activation units	ReLU		
optimizer	Adam		
optimizer learning rate	10^{-3}		

Table 2: Parameters used for Atari games.

Network Parameter	Value	Algorithmic Parameter	Value
T (total steps)	5×10^7	γ (discount rate)	0.99
C (interaction period)	4	$\tau_{\texttt{MVI}(q)}$ (MVI(q) entropy coefficient)	10
$\lvert B \rvert$ (buffer size)	1×10^6	$\alpha_{\texttt{MVI}(q)}$ (MVI(q) advantage coefficient)	0.9
B_t (batch size)	32	$\tau_{\texttt{Tsallis}}$ (Tsallis-VI entropy coef.)	10
I (update period)	8000	$\alpha_{\texttt{M-VI}}$ (M-VI advantage coefficient)	0.9
activation units	ReLU	$\tau_{\texttt{M-VI}}$ (M-VI entropy coefficient)	0.03
optimizer	Adam	ϵ (epsilon greedy threshold)	$1.0 \rightarrow 0.01\vert_{10\%}$
optimizer learning rate	10^{-4}		
Q-network architecture			
$\text{Conv}^4_{8,8}32$ - $\text{Conv}^2_{4,4}64$ - $\text{Conv}^1_{3,3}64$ - FC512 - FC			

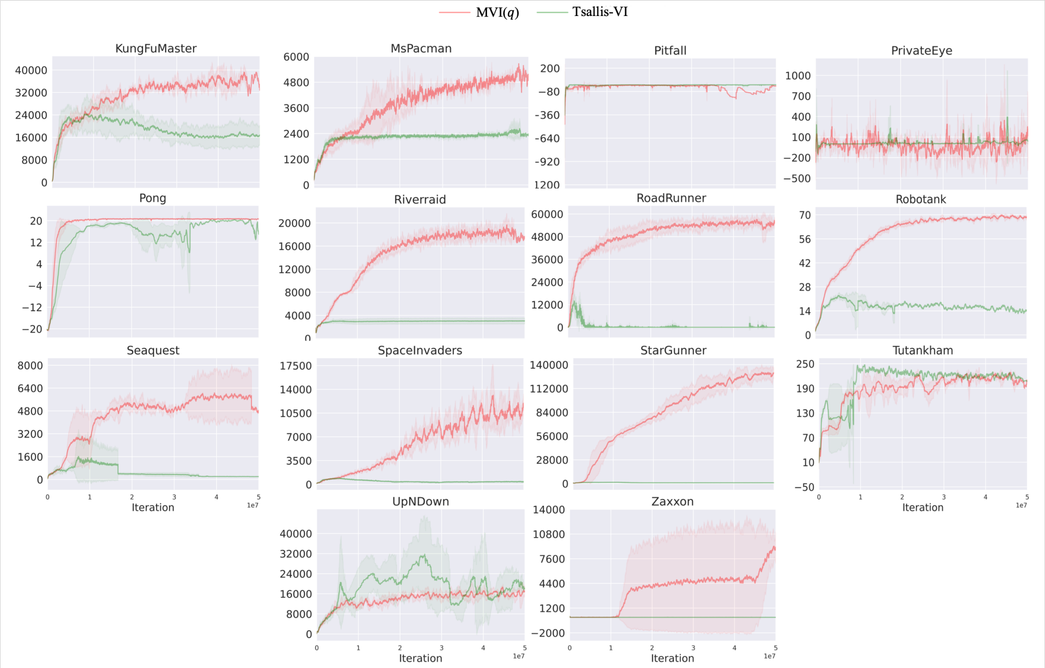

Figure 8: (cont'd) MVI(q) and Tsallis-VI on the selected Atari games.