Offline Reinforcement Learning with In-Sample Tsallis Regularized Policy

Department of Computing Science
Alberta University
Pittsburgh, PA 15213
hippo@cs.cranberry-lemon.edu

Abstract

Offline reinforcement learning methods learn from a fixed dataset collected by some behavior policy without further interaction with the environment. Such datasets often contain only a subset of the state and action spaces. Standard off-policy algorithms suffer from the extrapolation error known as the unrealistic action values for the actions not present in the dataset. The inaccurate value estimate can in turn cause out-of-distribution actions to be preferred in the policy improvement stage, leading to poor performance. Many methods propose to enforce the closedness between the learned policy and the behavior policy. In this paper, we propose to learn a in-sample Tsallis regularized policy that has support strictly within the dataset support. Our method also generalizes the in-sample softmax as a special case of the Tsallis policy where the support of the learned policy remains the same to the behavior policy. Moreover, if we assume the dataset is generated by Tsallis policies such that the absent actions are truncated, then we can guarantee the learned Tsallis policy is sufficiently similar to the behavior policy by showing bounded KL distance. This is the firs application of Tsallis regularization to offline RL to the best of our knowledge, where insufficient exploration as one of the main drawbacks of Tsallis regularized policies vanishes.

1 Introduction

2 Background

2.1 Reinforcement Learning

We focus on discrete-time discounted Markov Decision Processes (MDPs) expressed by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote state space and finite action space, respectively. $P(\cdot|s,a)$ denotes transition probability over the state space given state-action pair (s,a), and $r(s,a)$ defines the reward associated with that transition. $\gamma \in (0,1)$ is the discount factor. A policy $\pi(\cdot|s)$ is a mapping from the state space to distributions over actions. The state-action value function starting from (s,a) following policy π is defined as $Q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)|s_0 = s, a_0 = a\right]$. It is a classic result that there exists a stationary optimal policy that maximizes the cumulative return [19]. Its fixed point $Q_*(s,a)$ satisfies the Bellman optimality equation $Q_*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_{a'} Q_*(s',a')]$. Therefore, in practice it is common to update the action value function by acting greedily with respect to the current estimate.

We are interested in the MDPs that are *entropy-regularized*: the reward function is augmented with an additional entropic penality (or bonus). For example, the Shannon entropy $\mathcal{H}\left(\pi(\cdot|s)\right) := -\sum_a \pi(a|s) \ln \pi(a|s)$ is often added as a bonus to encourage the policy to be stochastic. The maximum Shannon entropy action value function satisfies $\tilde{Q}_\pi(s,a) = r(s,a) +$

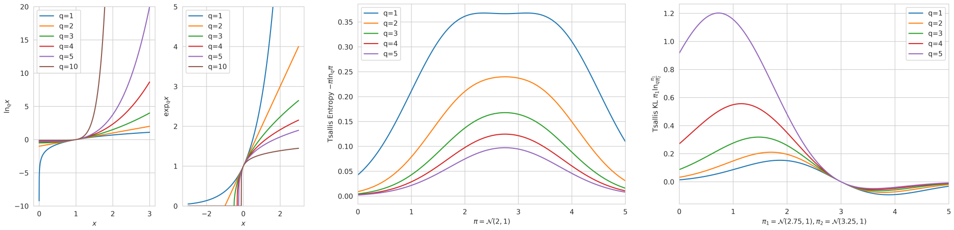

Figure 1: (Left) Behavior of q-logarithm and q-exponential functions. When $q = 1$ they respectively recover the standard logarithm and exponential. (Mid) Tsallis entropy of the Gaussian policy $\mathcal{N}(2, 1)$. (Right) Tsallis KL divergence between two Gaussian policies $\mathcal{N}(2.75, 1)$ and $\mathcal{N}(3.25, 1)$.

$\gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \mathbb{E}_{a \sim \pi(\cdot|s)} \left[\tilde{Q}_\pi(s, a) - \tau \mathcal{H}\left(\pi(\cdot|s)\right) \right]$, where τ denotes a coefficient. The maximum $r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\tau \ln \sum_a \exp\left(\tau^{-1} Q_\pi(s, a)\right) \right]$ is attained when the policy is the well-known Boltzmann softmax $\pi(a|s) \propto \exp\left(\tau^{-1} Q_\pi(s, a)\right)$, where $\propto$ denotes proportional to up to a constant not depending on actions. KL divergence $D_{KL}(\pi(\cdot|s) \,||\, \mu(\cdot|s)) := \sum_a \pi(a|s) \ln \frac{\pi(a|s)}{\mu(a|s)}$ is another popular choice, where μ is some baseline policy [2, 20, 24]. Unlike Shannon entropy, KL divergence is often added as a penalty to penalize large deviation from μ. The KL-regularized optimal policy takes the form $\pi(a|s) \propto \mu(a|s) \exp\left(\tau^{-1} Q_\pi(s, a)\right)$, where we overloaded the coefficient τ for Shannon entropy.

2.2 q-statistics and Tsallis regularization

In this paper we consider a broad class of less studied entropic regularizer known as the Tsallis entropy $S_q(\pi(\cdot|s)) := \frac{1}{q-1}\left(\sum_a \pi^q(a|s) - 1\right)$, where $q \in \mathbb{R}_+$. It is also called generalized entropy since it generalizes the Shannon entropy [22, 23]. Tsallis entropy can also be defined by the q-logarithm in a similar manner to the standard logarithm defining Shannon entropy, hence eases the notation and derivation. q-logarithm and its unique inverse function q-exponential are defined by [23]:

$$\ln_q x = \frac{x^{q-1} - 1}{q - 1}, \quad \exp_q x = \left[1 + (q-1)x\right]_+^{\frac{1}{q-1}}, \quad S_q(\pi(\cdot|s)) = -\sum_a \pi(a|s) \ln_q \pi(a|s), \tag{1}$$

where $[\cdot]_+ = \max\{\cdot, 0\}$. When $q \to 1$, the q-logarithm (resp. q-exponential) recovers the standard logarithm (resp. exponential) and hence Tsallis entropy degenerates to Shannon entropy. When $q = 2$, we arrive at the Tsallis sparse entropy $S_2(\pi(\cdot|s)) := \pi(a|s)\left(1 - \pi(a|s)\right)$ [5, 15]. The name sparse entropy comes from the fact that the regularizer leads to sparse support for the resulting policy [3, 17]. When $q = \infty$, the regularizer vanishes. In Figure 1 we plot the behavior of q-statistics under different q.

We can write out the regularized policy for general Tsallis entropy regularization for any q[1]:

$$\pi(a|s) = \exp_q\left(\frac{Q(s, a)}{\tau} - \psi\left(\frac{Q(s, \cdot)}{\tau}\right)\right), \; \psi\left(\frac{Q(s, \cdot)}{\tau}\right) \doteq \frac{\sum_{a \in K(s)} \frac{Q(s,a)}{\tau} - 1}{|K(s)|} + \frac{1}{q - 1}. \tag{2}$$

$K(s)$ is the set of highest-value actions satisfying $1 + i\frac{Q(s,a_{(i)})}{\tau} > \sum_{j=1}^{i} \frac{Q(s,a_{(j)})}{\tau}$, with $a_{(j)}$ denotes the j-th largest action. Intuitively, the policy first sorts actions by $a_{(1)}, \ldots, a_{(|\mathcal{A}|)}$ and then compute the threshold ψ. Suppose $Q(s, a_{(j+1)}) \le \psi \le Q(s, a_{(j)})$, then actions from $a_{(j+1)}, \ldots, a_{(|\mathcal{A}|)}$ are truncated and have zero probability of being selected. The truncation effect can be controlled by τ.

Note that for $q \neq 1, 2, \infty$ the policy does not have a closed-form solution. Nonetheless we can resort to the Taylor's expansion to obtain an approximate policy, see Appendix A for details. For those cases Eq. (2) still holds, but may need an additional renormalization step to guarantee a valid probability distribution. Therefore, in the rest of the paper, we write $\pi(a|s) \propto \exp_q\left(\frac{Q(s,a)}{\tau}\right)$. The effect of q lies in defining the range of the acceptable actions in $K(s)$ or extent of action truncation. Indeed, we

[1] Different to prior works [15, 5], our definition of ψ has an additional $\frac{1}{q-1}$ term to accommodate the q-exponential policy, see Appendix A for derivation.

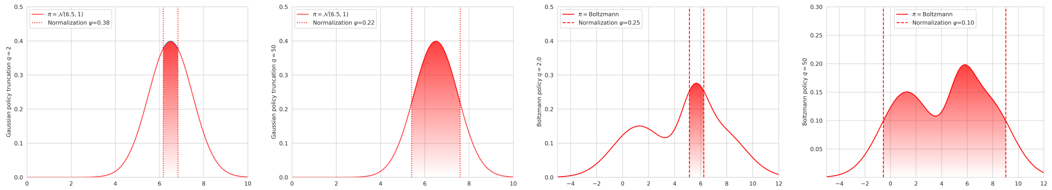

Figure 2: The sparsemax operator acting upon Gaussian and Boltzmann policies for $q = 2$ and $q = 50$ by truncating actions with values lower than ψ. As $q \rightarrow \infty$ the policy tends toward uniform distribution. We assume the The dataset is generated by Tsallis policies such that the actions absent from the offline dataset are those being truncated.

can show that as $q \rightarrow \infty$, the unnormalized Taylor's expansion for the policy $\pi(a|s) \rightarrow 1, \forall a$, that is, after normalization it becomes a uniform distribution. We plot the behavior of action truncation in Figure 2 for two commonly used policy classes, and indeed as q gets larger the truncation becomes weaker.

Very recently Zhu et al. [28] proposed in RL to exploit Tsallis KL divergence which generalizes Tsallis entropy. They proved that the regularized optimal policy takes a similar form to the optimal KL-regularized policy:

$$D^q_{KL}(\pi(\cdot|s) \,||\, \mu(\cdot|s)) = \sum_a \pi(a|s) \ln_q \frac{\pi(a|s)}{\mu(a|s)}, \quad \pi(a|s) = \mu(a|s) \exp_q \left(\frac{Q_\pi(s,a)}{\tau} - \psi\left(\frac{Q_\pi(s,a)}{\tau}\right)\right). \tag{3}$$

We can also resort to Taylor's expansion and renormalization to obtain approximate policies, hence the policy can be written as $\pi(a|s) \propto \mu(a|s) \exp_q \left(\frac{Q_\pi(s,a)}{\tau}\right)$. It is worth noting that unlike the KL divergence, the Tsallis KL divergence cannot be decomposed as $\ln_q \frac{\pi(a|s)}{\mu(a|s)} \neq \ln_q \pi(a|s) - \ln_q \mu(a|s)$.

2.3 Offline Reinforcement Learning

We consider the problem of offline RL, where the agent cannot interact with the environment and instead learn from a fixed dataset $\mathcal{D} = \{(s, a, r, s')_{1:N}\}$ collected by some unknown behavior policy $\pi_\mathcal{D}$. The dataset $\mathcal{D}$ typically contains only a small subset of the $\mathcal{S} \times \mathcal{A}$ space. Standard off-policy algorithms are known to suffer from the extrapolation error referring to the function approximator erroneously estimate action values for those out-of-distribution (OOD) actions not present in the dataset. Extrapolation error can happen as a result of e.g. the smoothness of neural networks when the update is around some low sample region [10]. In the policy improvement step, the unrealistic high values of the OOD actions lead to target values that are further in favor of sampling OOD actions, causing a vicious loop.

It is worth noting that unlike for online RL, where the OOD actions can lead to sampling more around the low sample region and eventually correction of the values, in offline RL this is impossible. Instead, many of the existing works propose to force the learned policy to be close to the behavior policy [6, 8, 7, 18]. This is often achieved by adding a regularizer in the policy update penalizing deviating from the behavior policy [9, 14, 11, 25], e.g. by incorporating KL divergence $\max_\pi \mathbb{E}_{s\sim\mathcal{D}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q(s,a)\right] - \tau D_{KL}(\pi(\cdot|s) \,||\, \pi_\mathcal{D}(\cdot|s))\right]$. The critic is typically updated by minimizing the loss $\mathbb{E}_{(s,a,s')\sim\mathcal{D}}\mathbb{E}_{a\sim\pi(\cdot|s)}\left[\left(r(s,a) + \gamma Q(s',a') - Q(s,a)\right)^2\right]$. The regularization leads to the policy form $\pi(a|s) \propto \pi_\mathcal{D}(a|s) \exp\left(\tau^{-1} Q_\pi(s,a)\right)$ where the learned policy must have the same support of the behavior policy. This idea is shared by the *in-sample* methods that enforce the support of the learned policy to be within the behavior policy $\pi \preceq \pi_\mathcal{D}$ [8, 13, 26], which is achieved by only updating for the actions in the dataset.

3 Method

3.1 In-Sample Softmax for Offline RL

To alleviate the OOD error, Fujimoto et al. [8] proposed the in-sample Bellman optimality equation to update only for actions present in the dataset:

$$Q_{*,\pi_{\mathcal{D}}}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' : \pi_{\mathcal{D}}(a'|s')>0} Q_{*,\pi_{\mathcal{D}}}(s',a') \right]. \tag{4}$$

Later, Xiao et al. [26] proposed in-sample softmax to better estimate the policy inside the bracket since in the continuous case the hard max operator might be difficult to solve for. By regularizing with the Shannon entropy $\tau \mathcal{H}(\pi)$ and imposing the dataset support constraint, the in-sample softmax Bellman optimality equation has the following evaluation step:

$$Q_{*,\pi_{\mathcal{D}}}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\tau \ln \sum_{a' : \pi_{\mathcal{D}}(a'|s')>0} \exp\left(\tau^{-1} Q_{*,\pi_{\mathcal{D}}}(s',a')\right) \right]. \tag{5}$$

As the dataset support constraint poses a challenge to implementation, Xiao et al. [26] proposed to transform the summation into an expectation to avoid directly computing the constraint:

$$\begin{aligned} \sum_{a' : \pi_{\mathcal{D}}(a'|s')>0} \exp\left(\tau^{-1} Q_{*,\pi_{\mathcal{D}}}(s',a')\right) &= \sum_{a' : \pi_{\mathcal{D}}(a'|s')>0} \frac{\pi_{\mathcal{D}}(a'|s')}{\pi_{\mathcal{D}}(a'|s')} \exp\left(\tau^{-1} Q_{*,\pi_{\mathcal{D}}}(s',a')\right) \\ &= \mathbb{E}_{a' \sim \pi_{\mathcal{D}}(\cdot|s')} \left[\exp\left(\tau^{-1} Q_{*,\pi_{\mathcal{D}}}(s',a') - \ln \pi_{\mathcal{D}}(a'|s')\right) \right]. \end{aligned} \tag{6}$$

The expectation can be approximated by Monte-Carlo sampling actions from the dataset. Since the term $\exp\left(\tau^{-1} Q_{*,\pi_{\mathcal{D}}}(s',a')\right)$ appears also in the regularized softmax policy, in-sample softmax updates the policy towards

$$\pi_{t+1,\pi_{\mathcal{D}}} \propto \pi_{\mathcal{D}}(a|s) \exp\left(\frac{Q_{t,\pi_{\mathcal{D}}}(s,a)}{\tau} - \ln \hat{\pi}_{\mathcal{D}}(a|s) \right), \tag{7}$$

where $\hat{\pi}_{\mathcal{D}}$ inside the exponential function is learned to imitate the behavior policy to avoid $\pi_{\mathcal{D}} = 0$ leading to an unbounded log-policy.

The benefits of the in-sample softmax can be interpreted as (1) softmax is easier to compute than hard max in the continuous action setting; (2) the in-sample softmax policy Eq. (7) can be seen as a KL-regularized policy and hence the dataset support constraint is satisfied. In fact, the support of $\pi_{\pi_{\mathcal{D}},k+1}$ should be exactly the same as $\pi_{\mathcal{D}}$ since $\exp(\frac{\cdot}{\tau}) > 0$ as long as $\tau \neq \infty$. We now explain why Eq. (7) can be regarded as a KL-regularized policy: it can be decomposed into two terms: the first term $\pi_{\mathcal{D}}(a|s) \exp\left(\frac{Q_{t,\pi_{\mathcal{D}}}(s,a)}{\tau}\right)$ acts as a KL-regularized policy with respect to the behavior policy; the second term $\exp\left(-\ln \hat{\pi}_{\mathcal{D}}(a|s)\right)$ can be seen as induced by another regularization $-\sum_a \pi(a|s) \ln \hat{\pi}_{\mathcal{D}}(a|s)$, which is the cross entropy between the in-sample softmax policy and the (learned) behavior policy.

However, the fact that the softmax policies always have full support indicates there is a persistent gap to Eq. (4). Since we want to improve upon the behavior policy, it is expected that action candidates should gradually narrow down to the maximizer. Furthermore, since Eq. (7) is only implicitly regularizing the policy and no explicit KL regularization for the value estimation [25], there is no guarantee that the learned policy should be close to the behavior policy.

3.2 In-Sample Tsallis Regularization

We propose to replace the Shannon entropy in in-sample softmax to the Tsallis entropy. The seemingly simple replacement, however, leads to drastically different behavior of the policy. Indeed, since we can control the truncation effect of the Tsallis regularized policies, we have the support of $\pi^{\text{Tsallis}}_{\pi_{\mathcal{D}},k+1} \preceq \pi_{\mathcal{D}}$, i.e. the support of the learned policy is either equal or within the dataset support.

Furthermore, Tsallis policies provide an means to formulate the common assumption that actions not present in the dataset are with low probability of being selected [12]. If we assume the behavior policy is also a Tsallis policy, then the support constraint $a : \pi_{\mathcal{D}}(a|s) > 0$ can be naturally replaced to the truncation criterion $a \in K(s)$; i.e.,

$$\pi_{\mathcal{D}}(a|s) \propto \exp_q\left(\frac{Q_{\pi_{\mathcal{D}}}(s,a)}{\tau_{\mathcal{D}}}\right), \qquad \sum_{a \in K_{\mathcal{D}}(s)} \pi_{\mathcal{D}}(a|s) = 1, \tag{8}$$

where $\tau_{\mathcal{D}}$ is an unknown coefficient and $K_{\mathcal{D}}(s)$ denotes the set of actions present in the dataset. Under the Tsallis behavior policy assumption, the dataset support constraint $a : \pi_{\mathcal{D}}(a|s) > 0$ coincides with the condition $a \in K_{\mathcal{D}}(s)$. Revisiting the in-sample hard-max Bellman equation Eq. (4):

$$Q_{*,\pi_{\mathcal{D}}}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\max_{a' \in K_{\mathcal{D}}(s)} Q_{*,\pi_{\mathcal{D}}}(s',a')\right].$$

However, the assumption of the Tsallis behavior policy alone is not useful. The power of the assumption manifests when we use also Tsallis regularized policy learning:

$$\pi_{t+1,\pi_{\mathcal{D}}}(a|s) \propto \exp_q\left(\frac{Q_{t,\pi_{\mathcal{D}}}(s,a)}{\tau}\right), \qquad \sum_{a \in K_{\mathcal{D}}(s)} \pi_{t+1,\pi_{\mathcal{D}}}(a|s) = 1. \tag{9}$$

$$Q_{t+1,\pi_{\mathcal{D}}}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\sum_{a' \in K_{\mathcal{D}}(s)} \pi_{t+1}(a'|s')\left(Q_{t,\pi_{\mathcal{D}}}(s',a') - \tau S_q(\pi_{t+1}(\cdot|s'))\right)\right]. \tag{10}$$

The term inside the bracket is known as q-maximum [16]. Similar to the softmax operators [1, 2], q-maximum is an approximation to the maximum with degree controlled by q.

More importantly, the support constraint is naturally satisfied since the learned Tsallis policy learns a new set of allowable actions $K_{t,q}$ from $K_{\mathcal{D}}(s)$ depending on q and iteration t. The set satisfies the condition $K_{t,q} \preceq K_{\mathcal{D}}$: i.e. $|K_{t,q}| \leq |K_{\mathcal{D}}|$ and support constraint $\pi_t \preceq \pi_{\mathcal{D}}$. Let us take $q = 2$ for example:

$$\begin{aligned}
Q_{t+1,\pi_{\mathcal{D}}}(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\sum_{a' \in K_{\mathcal{D}}(s)} \pi_{t+1}(a'|s')\left(Q_{t,\pi_{\mathcal{D}}}(s',a') - \tau S_2(\pi_{t+1}(\cdot|s'))\right)\right] \\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\tau \sum_{a' \in K_{\mathcal{D}}(s)} \left(\frac{Q_{t,\pi_{\mathcal{D}}}(s',a')}{\tau}\right)^2 - \left(\psi\left(\frac{Q_{t,\pi_{\mathcal{D}}}(s',a')}{\tau}\right) - 1\right)^2 + \tau\right] \\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\tau \sum_{a' \in K_{t,q}(s)} \left(\frac{Q_{t,\pi_{\mathcal{D}}}(s',a')}{\tau}\right)^2 - \left(\psi\left(\frac{Q_{t,\pi_{\mathcal{D}}}(s',a')}{\tau}\right) - 1\right)^2 + \tau\right].
\end{aligned} \tag{11}$$

The second equation is because $\max_\pi \sum_{a \in \pi(\cdot|s)} \pi(a|s)Q(s,a) - \tau S_2(\pi(\cdot|s))$ attains its maximum at $\tau \sum_{a \in K(s)} \left(\frac{Q(s,a)}{\tau}\right)^2 - \psi\left(\frac{Q(s,\cdot)}{\tau}\right)^2 + \tau$ [15, 17], and the last equation is due to $K_{t,q} \preceq K_{\mathcal{D}}$. Eq. (11) states that, one needs not to explicitly enforce the support constraint since it is automatically fulfilled by learning a new subset of allowable actions.

We now show that, if the Tsallis behavior policy assumption holds and we use Tsallis regularized policy learning, then the learned policy is guaranteed to stay close to the behavior policy in the sense that their KL divergence is uppper-bounded.

Theorem 1. *Suppose the dataset $\mathcal{D}$ is generated by a Tsallis behavior policy of entropic index q. Let $K_{t,q}(s) \preceq K_{\mathcal{D}}(s)$ denote the set of allowable actions at t-th iteration whose cardinality is smaller than $K_{\mathcal{D}}(s)$. Also let $\pi_t(a|s) \propto \exp_q\left(\frac{Q_{t-1}(s,a)}{\tau}\right)$ denote the learned policy. Then the KL divergence between π_t and the behavior policy can be upper bounded:*

$$D_{KL}(\pi_t(\cdot|s)\,||\,\pi_{\mathcal{D}}(\cdot|s)) \leq |K_{t,q}(s)|\left[\pi_t^q(a|s) - \pi_t^{q-1}(a|s) + \pi_t(a|s) - \frac{q-3}{q-1} + \pi_{\mathcal{D}}^{q-2}(a|s) - \pi_{\mathcal{D}}^{q-1}(a|s)\right]. \tag{12}$$

The bound is suggestive. For $q = 1$ (the in-sample softmax case) the bound is not useful and simply states the KL divergence may be unbounded. On the other hand, choosing any $q > 1$ brings an upper bound of at most $4|K_{t,q}(s)|$. When $q = 2$, the in-sample sparsemax has KL divergence to the behavior policy bounded by $|K_{t,q}(s)|\left(\pi_t(a|s) - \pi_{\mathcal{D}}(a|s) + 2\right)$. However, it should be noted that there is a trade-off between the power of policies π^q and the cardinality of $K_{t,q}(s)$: $K_{t,q}(s)$ tends to collect all actions when $q \to \infty$.

Proof. We first prove the following lemma:

Lemma 1. *The difference between the standard logarithm and q-logarithm can be expressed by:*

$$\ln x - \ln_q x = (q-1)\left[\frac{d}{dq}\ln_q x - \ln x \ln_q x\right].$$

Proof. Let us begin with the right hand side

$$\begin{aligned}
&(q-1)\left[\frac{d}{dq}\ln_q x - \ln x \ln_q x\right] = (q-1)\left[\frac{d}{dq}\frac{x^{q-1}-1}{q-1} - \ln x \ln_q x\right] \\
&= (q-1)\left[\frac{(x^{q-1}-1)'(q-1) - (x^{q-1}-1)(q-1)'}{(q-1)^2} - \ln x \ln_q x\right] \\
&= (q-1)\left[\frac{(q-1)x^{q-1}\ln x - (x^{q-1}-1)}{(q-1)^2} - \ln x \ln_q x\right] \\
&= x^{q-1}\ln x - \ln_q x - (q-1)\ln x \ln_q x = ((q-1)\ln_q x + 1)\ln x - \ln_q x - (q-1)\ln x \ln_q x \\
&= \ln x - \ln_q x.
\end{aligned}$$

□

With the lemma, we have the following theorem indicating the Tsallis backward learning policies have bounded distance to the behavior policy: We decompose the KL divergence into three terms and apply Lemma 1:

$$\begin{aligned}
&D_{KL}(\pi_t(\cdot|s)\,||\,\pi_{\mathcal{D}}(\cdot|s)) = \mathbb{E}_{a\sim\pi_t(\cdot|s)}\left[\ln \pi_t(a|s) - \ln \pi_{\mathcal{D}}(a|s)\right] \\
&= \mathbb{E}_{a\sim\pi_t(\cdot|s)}\left[\underbrace{\ln \pi_t(a|s) - \ln_q \pi_t(a|s)}_{(1)} + \underbrace{\ln_q \pi_t(a|s) - \ln_q \pi_{\mathcal{D}}(a|s)}_{(2)} + \underbrace{\ln_q \pi_{\mathcal{D}}(a|s) - \ln \pi_{\mathcal{D}}(a|s)}_{(3)}\right].
\end{aligned} \tag{13}$$

Let us now respectively bound the three terms:

$$\begin{aligned}
(1):\ln \pi_t(a|s) - \ln_q \pi_t(a|s) &= (q-1)\left[\frac{d}{dq}\ln_q \pi_t(a|s) - \ln_q \pi_t(a|s)\ln \pi_t(a|s)\right] \\
&= \pi_t^{q-1}(a|s)\ln \pi_t(a|s) - \ln_q \pi_t(a|s) - (q-1)\ln_q \pi_t(a|s)\ln \pi_t(a|s) \\
&\leq \pi_t^{q-1}(a|s)\ln \pi_t(a|s) + \frac{1}{q-1} + \ln \pi_t(a|s) \\
&\leq \left(\pi_t^{q}(a|s) - \pi_t^{q-1}(a|s)\right) + \pi_t(a|s) - \frac{q-2}{q-1},
\end{aligned} \tag{14}$$

where we leveraged $\ln x \leq x - 1$ and $\ln_q \exp_q(x) = x$ both when $x > 0$. Considering the definition of $\pi_t(a|s) \propto \exp_q\left(\frac{Q_{t-1}(s,a)}{\tau} - \psi\left(\frac{Q_{t-1}(s,\cdot)}{\tau}\right)\right)$ and $\exp_q(x) = [1 + (q-1)x]_+^{\frac{1}{q-1}}$, we must have $\pi_t(a|s) > 0 \Leftrightarrow a \in K_{t-1,q}(s) \Leftrightarrow -\frac{1}{q-1} \leq \frac{Q_{t-1}(s,a)}{\tau} - \psi\left(\frac{Q_{t-1}(s,\cdot)}{\tau}\right) \leq 0$. If $a \notin K_{t-1,q}(s)$, then $\ln \pi_t(a|s) = -\infty$ and the KL term is unbounded. The same fact is exploited to yield an upper bound $\frac{1}{q-1}$ for (2). We now work with (3):

$$\begin{aligned}
&(3):\ln_q \pi_{\mathcal{D}}(a|s) - \ln \pi_{\mathcal{D}}(a|s) = -(q-1)\left[\frac{d}{dq}\ln_q \pi_{\mathcal{D}}(a|s) - \ln_q \pi_{\mathcal{D}}(a|s)\ln \pi_{\mathcal{D}}(a|s)\right] \\
&\quad\leq -\pi_{\mathcal{D}}^{q-1}(a|s)\ln \pi_{\mathcal{D}}(a|s) \leq -\pi_{\mathcal{D}}^{q-1}(a|s)\left(1 - \frac{1}{\pi_{\mathcal{D}}(a|s)}\right) = \pi_{\mathcal{D}}^{q-2}(a|s) - \pi_{\mathcal{D}}^{q-1}(a|s).
\end{aligned} \tag{15}$$

Putting all terms together, we arrive at the upper bound that

$$
\begin{aligned}
D_{KL}(\pi_t(\cdot|s)\,||\,\pi_{\mathcal{D}}(\cdot|s)) &\leq \sum_{a\in K_{t,q}(s)} \pi_t(a|s)\left[\pi_t^q(a|s) - \pi_t^{q-1}(a|s) + \pi_t(a|s) - \frac{q-3}{q-1} + \pi_{\mathcal{D}}^{q-2}(a|s) - \pi_{\mathcal{D}}^{q-1}(a|s)\right] \\
&\leq K_{t,q}(s)\left[\pi_t^q(a|s) - \pi_t^{q-1}(a|s) + \pi_t(a|s) - \frac{q-3}{q-1} + \pi_{\mathcal{D}}^{q-2}(a|s) - \pi_{\mathcal{D}}^{q-1}(a|s)\right].
\end{aligned}
$$

□

The assumption of Tsallis behavior policy is not restrictive. If all actions present, then it corresponds to the case of $q = 1$ where the policy has full support. On the other hand, different levels of missingness can be simulated by both q and τ.

Another advantage of Eq. (10) is that for $q > 1$, the policy is a variant of categorical distributions which is less susceptible to numerical issues than exponential functions [21]. Furthermore, Eq. (10) can switch between greedy policy and multimodal policy. The former is achieved when all but one action have values lower than the threshold; while vice versa for the latter.

In-sample Tsallis Policy Estimation Similar to [26], we want to directly use the actions from the dataset to estimate our policy. We do a similar step to Eq. (7):

$$
\begin{aligned}
&\pi_{t+1,\pi_{\mathcal{D}}}(a|s) \propto \exp_q\left(\frac{1}{\tau}Q_{t,\pi_{\mathcal{D}}}(s,a)\right) = \pi_{\mathcal{D}}(a|s)\pi_{\mathcal{D}}(a|s)^{-1}\exp_q\left(\frac{1}{\tau}Q_{t,\pi_{\mathcal{D}}}(s,a)\right) \\
&= \pi_{\mathcal{D}}(a|s)\exp_q\left(\ln_q\frac{1}{\pi_{\mathcal{D}}(a|s)}\right)\exp_q\left(\frac{1}{\tau}Q_{t,\pi_{\mathcal{D}}}(s,a)\right) \\
&= \pi_{\mathcal{D}}(a|s)\left(\exp_q\left(\frac{1}{\tau}Q_{t,\pi_{\mathcal{D}}}(s,a) + \ln_q\frac{1}{\pi_{\mathcal{D}}(a|s)}\right)^{q-1} - (q-1)^2\frac{1}{\tau}Q_{t,\pi_{\mathcal{D}}}(s,a)\ln_q\frac{1}{\pi_{\mathcal{D}}(a|s)}\right)^{\frac{1}{q-1}}.
\end{aligned}
\tag{16}
$$

In the last step we made use of the relationship $\left(\exp_q x \cdot \exp_q y\right)^{q-1} = \exp_q(x+y)^{q-1} + (q-1)^2 xy$ [27].

Remark. *Tsallis entropy regularization has not been popular since its proposal in RL [15]. One of the main reasons is the sparsemax policies are not suitable for online RL since the exploration is handicapped resulted from the action truncation. However, in offline RL this drawback vanishes, and theoretically it allows for bounding the distance to the behavior policy, guaranteeing less OOD actions.*

4 Implementation

Let θ, ϕ, ω denote the parametrization of networks for Q, $\pi_{\pi_{\mathcal{D}}}$, $\pi_{\mathcal{D}}$, respectively. In-sample softmax updates the policy towards

$$
\pi_{\pi_{\mathcal{D}},Q_\theta}(a|s) = \pi_{\mathcal{D}}(a|s)\exp\left(\frac{Q_\theta(s,a) - Z(s)}{\tau} - \ln\pi_\omega(a|s)\right), \tag{17}
$$

where $Z(s)$ denotes the normalization constant and is necessary since the policy is updated by minimizing KL divergence:

$$
\begin{aligned}
D_{KL}(\pi_{\pi_{\mathcal{D}},Q_\theta}(\cdot|s)\,||\,\pi_\phi(\cdot|s)) &= \mathbb{E}_{a\sim\pi_{\pi_{\mathcal{D}},Q_\theta}(\cdot|s)}\left[\ln\pi_{\pi_{\mathcal{D}},Q_\theta}(a|s) - \ln\pi_\phi(a|s)\right] \\
&= \mathbb{E}_{a\sim\pi_{\mathcal{D}}(\cdot|s)}\left[-\exp\left(\frac{Q_\theta(s,a) - Z(s)}{\tau} - \ln\pi_\omega(a|s)\right)\ln\pi_\phi(a|s)\right],
\end{aligned}
$$

where the $\pi_{\mathcal{D}}$ term in $\pi_{\pi_{\mathcal{D}},Q_\theta}$ is absorbed into the expectation, so the KL divergence loss can be minimized by sampling actions from the offline dataset.

We follow the same setup here, but replacing every appearance of $\ln, \exp$ to their q-logarithm and q-exponential counterpart. Similar to the discussion after Eq. (7), the Tsallis policy we derived here can be seen as the result from Tsallis KL regularization, plus another regularization that gives rise to the additional term inside the $\exp_q$ function. In the implementation, we choose the sparsemax parametrization $q = 2$, which gives the following Tsallis in-sample sparsemax actor-critic update rule:

$$\mathcal{L}_{\text{actor}}(\phi) = -\mathbb{E}_{s,a\sim\mathcal{D}}\left[\left(\exp_2\left(\frac{1}{\tau}Q_\theta(s,a) + \ln_2\frac{1}{\pi_\omega(a|s)}\right) - \frac{1}{\tau}Q_\theta(s,a)\ln_q\frac{1}{\pi_\omega(a|s)}\right)\ln\pi_\phi(a|s)\right], \tag{18}$$

$$\mathcal{L}_{\text{baseline}}(\zeta) = \mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\phi(s)}\left[\left(v_\zeta(s) - Q_\theta(s,a) - \tau\ln_2\pi_\phi(a|s)\right)^2\right], \tag{19}$$

$$\mathcal{L}_{\text{critic}}(\theta) = \mathbb{E}_{s,a,r,s'\sim\mathcal{D}}\left[\left(r + \gamma v_\zeta(s') - Q_\theta(s,a)\right)^2\right]. \tag{20}$$

References

[1] K. Asadi and M. L. Littman. An alternative softmax operator for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 243–252, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[2] M. G. Azar, V. Gómez, and H. J. Kappen. Dynamic policy programming. *Journal of Machine Learning Research*, 13(1):3207–3245, 2012.

[3] M. Blondel, A. F. Martins, and V. Niculae. Learning with fenchel-young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020.

[4] G. Chen, Y. Peng, and M. Zhang. Effective exploration for deep reinforcement learning via bootstrapped q-ensembles under tsallis entropy regularization. *arXiv:abs/1809.00403*, 2018. URL `http://arxiv.org/abs/1809.00403`.

[5] Y. Chow, O. Nachum, and M. Ghavamzadeh. Path consistency learning in Tsallis entropy regularized MDPs. In *International Conference on Machine Learning*, pages 979–988, 2018.

[6] R. Dadashi, S. Rezaeifar, N. Vieillard, L. Hussenot, O. Pietquin, and M. Geist. Offline reinforcement learning with pseudometric learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2307–2318, 2021.

[7] S. Fujimoto and S. S. Gu. A minimalist approach to offline reinforcement learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[8] S. Fujimoto, D. Meger, and D. Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2052–2062, 2019.

[9] S. K. S. Ghasemipour, D. Schuurmans, and S. S. Gu. Emaq: Expected-max q-learning operator for simple yet effective offline and online rl. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 3682–3691, 2021.

[10] C. Gulcehre, S. G. Colmenarejo, Z. Wang, J. Sygnowski, T. Paine, K. Zolna, Y. Chen, M. Hoffman, R. Pascanu, and N. de Freitas. Regularized behavior value estimation, 2021.

[11] N. Jaques, A. Ghandeharioun, J. H. Shen, C. Ferguson, A. Lapedriza, N. Jones, S. Gu, and R. Picard. Way off-policy batch deep reinforcement learning of human preferences in dialog, 2020. URL `https://openreview.net/forum?id=rJl5rRVFvH`.

[12] I. Kostrikov, R. Fergus, J. Tompson, and O. Nachum. Offline reinforcement learning with fisher divergence critic regularization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5774–5783, 2021.

[13] I. Kostrikov, A. Nair, and S. Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=68n2s9ZJWF8`.

[14] A. Kumar, J. Fu, G. Tucker, and S. Levine. *Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction*. 2019.

[15] K. Lee, S. Choi, and S. Oh. Sparse markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3:1466–1473, 2018.

[16] K. Lee, S. Kim, S. Lim, S. Choi, M. Hong, J. I. Kim, Y. Park, and S. Oh. Generalized tsallis entropy reinforcement learning and its application to soft mobile robots. In *Robotics: Science and Systems XVI*, pages 1–10, 2020.

[17] A. F. T. Martins and R. F. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on Machine Learning*, page 1614–1623, 2016.

[18] A. Nair, M. Dalal, A. Gupta, and S. Levine. {AWAC}: Accelerating online reinforcement learning with offline datasets, 2021. URL `https://openreview.net/forum?id=OJiM1R3jAtZ`.

[19] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[20] K. Rawlik, M. Toussaint, and S. Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference (extended abstract). In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, page 3052–3056, 2013.

[21] Y.-H. H. Tsai, M. Q. Ma, M. Yang, H. Zhao, L.-P. Morency, and R. Salakhutdinov. Self-supervised representation learning with relative predictive coding. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=068E_JSq9O`.

[22] C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

[23] C. Tsallis. *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*. Springer New York, 2009. ISBN 9780387853581.

[24] N. Vieillard, T. Kozuno, B. Scherrer, O. Pietquin, R. Munos, and M. Geist. Leverage the average: an analysis of regularization in rl. In *Advances in Neural Information Processing Systems 33*, pages 1–12, 2020.

[25] Y. Wu, G. Tucker, and O. Nachum. Behavior regularized offline reinforcement learning, 2020. URL `https://openreview.net/forum?id=BJg9hTNKPH`.

[26] C. Xiao, H. Wang, Y. Pan, A. White, and M. White. The in-sample softmax for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=u-RuvyDYqCM`.

[27] T. Yamano. Some properties of q-logarithm and q-exponential functions in tsallis statistics. *Physica A: Statistical Mechanics and its Applications*, 305(3):486–496, 2002.

[28] L. Zhu, Z. Chen, T. Matsubara, and M. White. Generalized munchausen reinforcement learning using tsallis kl divergence, 2023.

A Derivation

It is worth noting that, for $q \neq 1$ the regularized policy is given by

$$\pi^*(a|s) = \sqrt[q-1]{\left[\frac{Q_\pi(s,a)}{q\tau} - \psi\left(\frac{Q_\pi(s,\cdot)}{\tau}\right)\right]_+ (q-1)}, \tag{21}$$

the normalization function ψ can only be analytically solved when $q = 2, \infty$. When $q \neq 1, 2$, the closed-form expression of π, ψ might not exist. Following [4], we leverage the first order Taylor expansion $f(z) + f'(z)(x - z)$ on the policy Eq. (21), where we let $z = 1$, $x = \left[\frac{Q_\pi(s,a)}{q\tau} - \psi\left(\frac{Q_\pi(s,\cdot)}{q\tau}\right)\right]_+ \frac{q-1}{p}$, $f(x) = x^{\frac{1}{q-1}}$, $f'(x) = \frac{1}{q-1}x^{\frac{2-q}{q-1}}$. So that

$$\begin{aligned} \tilde{\pi}^*(a|s) &\approx f(z) + f'(z)(x - z) \\ &= 1 + \frac{1}{q-1}\left(\left(\frac{Q_\pi(s,a)}{q\tau} - \tilde{\psi}\left(\frac{Q_\pi(s,\cdot)}{q\tau}\right)\right)\frac{q-1}{p} - 1\right). \end{aligned} \tag{22}$$