Offline Reinforcement Learning with Tsallis Regularization

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

Abstract

We show that the Tsallis regularization is well suited to offline reinforcement learning, where its main drawback insufficient exploration vanishes. Furthermore, Tsallis entropy induces sparsemax policies which have a natural interpretation of regarding actions not present in the dataset as the truncated actions, hence no support constraint as is done in In-Sample softmax is necessary.

1 Introduction

2 Background

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. $d(\cdot)$ denotes the initial state distribution. $P(\cdot|s,a)$ denotes transition probability over the state space given state-action pair (s,a), and $r(s,a)$ defines the reward associated with that transition. When time step t is of concern, we write $r_t := r(s_t, a_t)$. $\gamma \in (0,1)$ is the discount factor. A policy $\pi(\cdot|s)$ is a mapping from the state space to distributions over actions. We define the state value function at a specific state s following policy π as $V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} r_t | s\right]$. Likewise, we define the state-action value function given initial action a_0 as $Q_\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} r_t | s_0 = s, a_0 = a\right]$, where the expectation is with respect to the policy π and transition probability P.

A standard approach to find the optimal value function Q_* is value iteration. To define the formulas for value iteration, it will be convenient to write these functions as vectors, $V_\pi \in \mathbb{R}^{|\mathcal{S}|}$ and $Q_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. For notational convenience, we define the inner product for any two functions $F_1, F_2 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ as $\langle F_1, F_2 \rangle \in \mathbb{R}^{|\mathcal{S}|}$, where we only take the inner product over actions. The Bellman operator acting upon any function $Q \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ can be defined as: $T_\pi Q := r + \gamma P_\pi Q$, where $P_\pi Q \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} := \mathbb{E}_{s'\sim P(\cdot|s,a), a'\sim\pi(\cdot|s')}[Q(s',a')] = P\langle \pi, Q\rangle$. When π is greedy with respect to Q, we have the Bellman optimality operator defined by $T_* Q := r + \gamma P_* Q$, $P_* Q = \mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_{a'} Q(s',a')]$. The above definitions should be understood as component-wise. Repeatedly applying the Bellman operator $T_\pi Q$ converges to the unique fixed point Q_π, and $T_* Q$ it converges to the optimal action value function $Q_* := Q_{\pi^*}$. This basic recursion can be modified with the addition of a regularizer $\Omega(\pi)$:

$$\begin{cases} \pi_{k+1} = \arg\max_\pi \langle \pi, Q_k \rangle, \\ Q_{k+1} = \left(T_{\pi_{k+1}} Q_k\right)^m, \end{cases} \qquad \begin{cases} \pi_{k+1} = \arg\max_\pi \left(\langle \pi, Q_k \rangle - \tau\Omega(\pi)\right), \\ Q_{k+1} = \left(T_{\pi_{k+1},\Omega} Q_k\right)^m \end{cases} \tag{1}$$

where $T_{\pi_{k+1},\Omega} Q_k = r + \gamma P(\langle \pi_{k+1}, Q_k \rangle - \tau\Omega(\pi_{k+1}))$ is the regularized Bellman operator [5]. This modified recursion is guaranteed to converge if Ω is concave in π. For standard (Shannon) entropy regularization, we use $\Omega(\pi) = -\mathcal{H}(\pi) = \langle \pi, \ln \pi \rangle$. The resulting optimal policy has $\pi_{k+1} \propto \exp\left(\tau^{-1} Q_k\right)$, where $\propto$ indicates *proportional to* up to a constant not depending on actions.

Preprint. Work in progress.

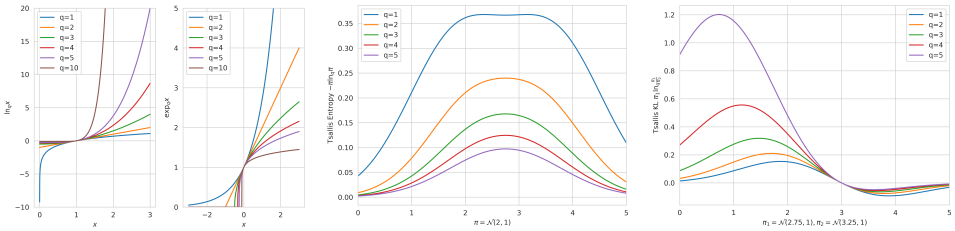

Figure 1: (Left) Behavior of q-logarithm and q-exponential functions. When $q = 1$ they respectively recover the standard logarithm and exponential. (Mid) Tsallis entropy of the Gaussian policy $\mathcal{N}(2, 1)$. (Right) Tsallis KL divergence between two Gaussian policies $\mathcal{N}(2.75, 1)$ and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

Another popular choice is KL divergence $\Omega(\pi) = D_{KL}(\pi \,||\, \mu) = \langle \pi, \ln \pi - \ln \mu \rangle$, which is more general since we recover Shannon entropy when we choose μ to be a uniform distribution, i.e. $\frac{1}{|\mathcal{A}|}$. In this work, when we say KL regularization, we mean the standard choice of setting $\mu = \pi_k$, the estimate from the previous update. Therefore, D_{KL} serves as a penalty to penalize aggressive policy changes. The optimal policy in this case takes the form $\pi_{k+1} \propto \pi_k \exp\left(\tau^{-1} Q_k\right)$. By induction, we can show this KL-regularized optimal policy π_{k+1} is a softmax over a uniform average over the history of action value estimates [16]:

$$\pi_{k+1} \propto \pi_k \exp\left(\tau^{-1} Q_k\right) \propto \cdots \propto \exp\left(\tau^{-1} \sum_{j=1}^{k} Q_j\right). \tag{2}$$

Using KL regularization has been shown to be theoretically superior to entropy regularization, in terms of error tolerance [1, 16, 8, 3].

The definitions of $\mathcal{H}(\cdot)$ and $D_{\text{KL}}(\cdot||\cdot)$ rely on the standard logarithm and its inverse (the exponential) and both induce softmax policies as an exponential over (weighted) action-values [6, 12]. Convergence properties of the resulting regularized algorithms have been well studied [7, 5, 16]. In this paper, we investigate Tsallis entropy and Tsallis KL divergence as the regularizer, which generalize Shannon entropy and KL divergence respectively.

3 q-statistics and Tsallis regularization

A more general form of entropy known as Tsallis entropy is proposed by [14], defined based on the q-logarithm that is referred to as *deformed logarithm* [15]. Let us revisit the definition of q-logarithm its unique inverse function q-exponential (analogous to the exponential function), see Figure 1. We can define the Tsallis entropy $S_q(\pi)$ in a similar way to Shannon entropy with q-logarithm :

$$q \in \mathbb{R}, \quad \ln_q x = \frac{x^{q-1} - 1}{q - 1}, \quad \exp_q x = \left[1 + (q-1)x\right]_+^{\frac{1}{q-1}}, \quad S_q(\pi) = -\left\langle \pi, \ln_q \pi \right\rangle. \tag{3}$$

When $q = 2$, we recover the Tsallis sparse entropy [11], which achieves the sparsity by projecting onto the probability simplex [2]. In RL, it has been investigated by [9, 10] and it is known that the optimal regularized policy takes the form:

$$\pi_{k+1}(a|s) = \left[\frac{Q_k(s,a)}{\tau} - \psi\left(\frac{Q_k(s,\cdot)}{\tau}\right)\right]_+ = \exp_2\left(\frac{Q_k(s,a)}{\tau} - \tilde{\psi}\left(\frac{Q_k(s,\cdot)}{\tau}\right)\right),$$

$$\psi\left(\frac{Q_k(s,\cdot)}{\tau}\right) \doteq \frac{\sum_{a \in K(s)} \frac{Q(s,a)}{\tau} - 1}{|K(s)|}, \; \tilde{\psi}\left(\frac{Q_k(s,\cdot)}{\tau}\right) := \psi\left(\frac{Q_k(s,\cdot)}{\tau}\right) + \frac{1}{q-1}$$

where $[\cdot]_+ = \max\{\cdot, 0\}$, $K(s)$ 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. (say something here about the general q cases, and how $q \to \infty$ the policy tends to uniform distribution, provide proof.)

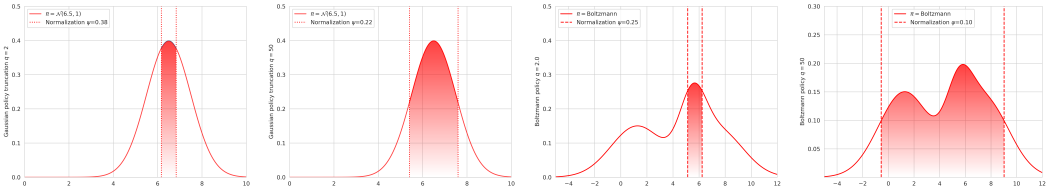

Figure 2: The sparsemax operator acting upon Gaussian and Boltzmann policies for $q = 2$ and $q = 50$ by truncating actions with values lower than ψ. As $q \to \infty$ the policy tends toward uniform distribution. We assume the The dataset is generated by Tsallis policies such that the actions absent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

Very recently Zhu et al. [19] proposed in RL to exploit Tsallis KL divergence which generalizes Tsallis entropy. They proved that the regularized optimal policy takes a similar form to Eq. (2):

$$D_{KL}^{q}(\pi \,||\, \mu) = \left\langle \pi, \ln_q \frac{\pi}{\mu} \right\rangle, \quad \pi_{k+1} = \pi_k \exp_q \left(\frac{Q_k}{\tau} - \psi \left(\frac{Q_k}{\tau} \right) \right). \tag{4}$$

Eq. (4) is important to our application of offline RL where we show the in-sample softmax becomes in-sample sparsemax.

4 In-Sample Softmax for Offline RL

To alleviate the out-of-distribution error, Fujimoto et al. [4] proposed the in-sample Bellman optimality equation to update only for actions present in the dataset:

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

Later, Xiao et al. [17] 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 adding negative entropy regularization $\Omega(\pi) = \tau \mathcal{H}(\pi)$ and imposing the data support constraint, the in-sample softmax Bellman optimality equation has the following evaluation step:

$$Q_{*,\pi_\mathcal{D}} = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\sum_{a': \pi_\mathcal{D}(a'|s') > 0} \pi(a'|s') \left(Q_{*,\pi_\mathcal{D}}(s', a') - \tau \mathcal{H}(\pi)(\cdot|s') \right) \right]. \tag{6}$$

By choosing the maximizer policy the above equation takes the form:

$$Q_{*,\pi_\mathcal{D}} = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\tau \ln \sum_{a': \pi_\mathcal{D}(a'|s') > 0} \exp\left(\tau^{-1} Q_{*,\pi_\mathcal{D}}(s', a') \right) \right]. \tag{7}$$

As the dataset support constraint poses a challenge to implementation, Xiao et al. [17] proposed to transform the summation into an expectation to avoid directly computing the constraint:

$$\begin{aligned} \sum_{a': \pi_\mathcal{D}(a'|s') > 0} \exp\left(\tau^{-1} Q_{*,\pi_\mathcal{D}}(s', a') \right) &= \sum_{a': \pi_\mathcal{D}(a'|s') > 0} \frac{\pi_\mathcal{D}(a'|s')}{\pi_\mathcal{D}(a'|s')} \exp\left(\tau^{-1} Q_{*,\pi_\mathcal{D}}(s', a') \right) \\ &= \mathbb{E}_{a' \sim \pi_\mathcal{D}(\cdot|s')} \left[\exp\left(\tau^{-1} Q_{*,\pi_\mathcal{D}}(s', a') - \ln \pi_\mathcal{D}(a'|s') \right) \right]. \end{aligned} \tag{8}$$

The expectation is 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

$$\pi_{\pi_\mathcal{D},k+1} \propto \pi_\mathcal{D}(a|s) \exp\left(\frac{Q_{\pi_\mathcal{D},k}(s,a)}{\tau} - \ln \hat{\pi}_\mathcal{D}(a|s) \right), \tag{9}$$

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.

We interpret Eq. (9) from another perspective. the in-sample softmax policy can be decomposed into two terms: the first term $\pi_{\mathcal{D}}(a|s)\exp\left(\frac{Q_{\pi_{\mathcal{D}},k}(s,a)}{\tau}\right)$ acts as a KL-regularized policy with respect to the behavior policy (see Eq. (2)); the second term $\exp\left(-\ln\hat{\pi}_{\mathcal{D}}(a|s)\right)$ can be seen as induced by another regularization $\Omega(\pi) = -\langle\pi, \ln\hat{\pi}_{\mathcal{D}}\rangle$, the cross entropy between the in-sample softmax policy and the behavior policy.

5 Forward-Backward Tsallis Learning for Offline RL

Summary:

- Forward: Tsallis sparsemax policies generate the dataset. This assumption is not very strong since Tsallis entropy generalizes Shannon entropy and the behavior policy class is sufficiently rich.
- Backward: learn the optimal policy with the modified Tsallis entropy regularization. The learned policy differs from the conventional sparsemax policies in that no thresholding is required.

The key to both in-sample max [4] and in-sample softmax [17] is the dataset support constraint. In this paper, we propose to exploit the properties of sparsemax policies to naturally satisfy this constraint. In a nutshell, we assume the behavior policies are Tsallis sparsemax policies, and therefore the actions not present in the dataset are with low probability and are exactly these truncated actions. For simplicity, suppose the dataset is generated by the Tsallis sparsemax policy:

$$\pi_{\mathcal{D}}(a|s) = \left[\frac{Q_{\pi_{\mathcal{D}}}(s,a)}{\tau_{\mathcal{D}}} - \psi\left(\frac{Q_{\pi_{\mathcal{D}}}(s,\cdot)}{\tau_{\mathcal{D}}}\right)\right]_+, \quad \sum_{a\in K_{\mathcal{D}}(s)}\pi_{\mathcal{D}}(a|s) = 1, \tag{10}$$

where $\tau_{\mathcal{D}}$ is a unknown coefficient and $K_{\mathcal{D}}(s)$ denotes the set of allowable actions. Now let us inspect the Bellman equation for the behavior policy:

$$\begin{aligned}
Q_{\pi_{\mathcal{D}}}(s,a) &= r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\sum_{a'}\pi_{\mathcal{D}}(a'|s')\left(Q_{\pi_{\mathcal{D}}}(s',a') - \tau_{\mathcal{D}}S_2(\pi_{\mathcal{D}})(\cdot|s')\right)\right]. \\
&= r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\tau_{\mathcal{D}}\left(\sum_{a'\in K_{\mathcal{D}}(s')}\left(\frac{Q_{\pi_{\mathcal{D}}}(s',a')}{\tau_{\mathcal{D}}}\right)^2 - \psi\left(\frac{Q_{\pi_{\mathcal{D}}}(s',\cdot)}{\tau_{\mathcal{D}}}\right)^2 + 1\right)\right] \\
&= r(s,a) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[\tau_{\mathcal{D}}\left(\sum_{a':\pi_{\mathcal{D}}(a'|s')>0}\left(\frac{Q_{\pi_{\mathcal{D}}}(s',a')}{\tau_{\mathcal{D}}}\right)^2 - \psi\left(\frac{Q_{\pi_{\mathcal{D}}}(s',\cdot)}{\tau_{\mathcal{D}}}\right)^2 + 1\right)\right].
\end{aligned} \tag{11}$$

Here, $\max_{\pi\in\Delta(\mathcal{A})}\pi(a|s)Q(s,a) + \tau S_2(\pi(\cdot|s))$ attains its maximum at $V(s) = \tau\sum_{a\in K(s)}\left(\frac{Q(s,a)}{\tau}\right)^2 - \psi\left(\frac{Q(s,\cdot)}{\tau}\right)^2 + \tau$ [9, 11]. It is also possible to assume that $\pi_{\mathcal{D}}(a|s) = \exp_q\left(\frac{Q_{\pi_{\mathcal{D}}}(s,a)}{\tau_{\mathcal{D}}} - \tilde{\psi}\left(\frac{Q_{\pi_{\mathcal{D}}}(s,\cdot)}{\tau_{\mathcal{D}}}\right)\right)$ for general $q > 2$, though in these cases the policy does not have a closed-form expression we have to resort to approximation. Under the Tsallis behavior policy assumption, the dataset support constraint $a : \pi_{\mathcal{D}}(a|s) > 0$ coincides with the condition $a \in K_{\mathcal{D}}(s)$. Moreover, the assumption actually assumes a forward process of learning a set of allowable actions $K_{\mathcal{D}}(s)$ given $Q_{\pi_{\mathcal{D}}}(s,a), \forall s,a$; and it naturally raises the possibility of *backward Tsallis learning*: given the set of allowable actions $K_{\mathcal{D}}(s)$, learn an optimal policy that resembles $\pi_{\mathcal{D}}$ in the sense of $\pi_* \preceq \pi_{\mathcal{D}}$, i.e. putting probability mass only on the actions in $K_{\mathcal{D}}$.

The backward Tsallis learning consists in learning $\pi_{t,\pi_{\mathcal{D}}} \preceq \pi_{\mathcal{D}}$ given $K_{\mathcal{D}}$.

$$\pi_{t+1,\pi_{\mathcal{D}}}(a|s) = \exp_q\left(\frac{Q_{t,\pi_{\mathcal{D}}}(s,a)}{\tau} - \tilde{\psi}\left(\frac{Q_{t,\pi_{\mathcal{D}}}(s,\cdot)}{\tau}\right)\right), \quad \sum_{a\in K_{\mathcal{D}}(s)}\pi_{t+1,\pi_{\mathcal{D}}} = 1. \tag{12}$$

The thresholding function together with the given $K_{\mathcal{D}}$ implies during the backward learning the new set of allowable actions $K_{t,q}$, where we make clear the dependence on q and iteration t. The set

should satisfy the condition $K_{t,q} \preceq K_{\mathcal{D}}$: i.e. the cardinality $|K_{t,q}| \leq |K_{\mathcal{D}}|$ and support constraint $\pi_t \preceq \pi_{\mathcal{D}}$. Another advantage lies in that for $q > 1$, the policy is a variant of categorical distributions which is less susceptible to numerical issues than exponential functions [13].

Why use Tsallis backward learning? If the Tsallis behavior policy assumption holds, and Tsallis backward learning is adopted, then the KL divergence between the learned policy and the behavior policy can be uppper-bounded. That being said, the learned policy is guaranteed to better respect the behavior policy.

Theorem 1. *Suppose the dataset $\mathcal{D}$ is generated by a Tsallis behavior policy of any entropic index q^* (it is not important). 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 the π_t and the behavior policy can be upper bounded:*

$$D_{KL}(\pi_t(\cdot|s) \,||\, \pi_{\mathcal{D}}(\cdot|s)) \leq |K_t(s)| \left[\pi_t^q(a|s) - \pi_t^{q-1}(a|s) + \pi_t(a|s) - \frac{q-3}{q-1} + \pi_{\mathcal{D}}^{q-2}(a|s) - \pi_{\mathcal{D}}^{q-1}(a|s)\right]. \tag{13}$$

The bound is suggesting. It explains why using in-sample softmax $q = 1$ may not be a good choice as in this case the KL divergence to the behavior policy is unbounded. On the other hand, choosing $q = 2$ which recovers sparsemax is 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. Before we state the main result, we 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:

$$\begin{aligned}
&D_{KL}(\pi_t(\cdot|s) \,||\, \pi_{\mathcal{D}}(\cdot|s)) = \mathbb{E}_{a\sim\pi_t(\cdot|s)}\left[\ln \pi_t(a|s) - \ln \pi_{\mathcal{D}}(a|s)\right] \\
&= \mathbb{E}_{a\sim\pi_t(\cdot|s)}\left[\underbrace{\ln \pi_t(a|s) - \ln_q \pi_t(a|s)}_{(1)} + \underbrace{\ln_q \pi_t(a|s) - \ln_q \pi_{\mathcal{D}}(a|s)}_{(2)} + \underbrace{\ln_q \pi_{\mathcal{D}}(a|s) - \ln \pi_{\mathcal{D}}(a|s)}_{(3)}\right].
\end{aligned} \tag{14}$$

Let us now respectively bound the three terms:

$$\begin{aligned}(1) : \ln \pi_t(a|s) - \ln_q \pi_t(a|s) &= (q-1)\left[\frac{d}{dq}\ln_q \pi_t(a|s) - \ln_q \pi_t(a|s)\ln \pi_t(a|s)\right] \\ &= \pi_t^{q-1}(a|s)\ln \pi_t(a|s) - \ln_q \pi_t(a|s) - (q-1)\ln_q \pi_t(a|s)\ln \pi_t(a|s) \\ &\leq \pi_t^{q-1}(a|s)\ln \pi_t(a|s) + \frac{1}{q-1} + \ln \pi_t(a|s) \\ &\leq \left(\pi_t^q(a|s) - \pi_t^{q-1}(a|s)\right) + \pi_t(a|s) - \frac{q-2}{q-1},\end{aligned} \tag{15}$$

where we leveraged $\ln x \leq x-1$ and the fact that $\forall a \in K(s),\ -\frac{1}{q-1} \leq \frac{Q_{t-1}(s,a)}{\tau} - \tilde{\psi}\left(\frac{Q_{t-1}(s,\cdot)}{\tau}\right) \leq 0$. If $a \notin K(s)$, 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):

$$\begin{aligned}(3) : \ln_q \pi_{\mathcal{D}}(a|s) - \ln \pi_{\mathcal{D}}(a|s) &= -(q-1)\left[\frac{d}{dq}\ln_q \pi_{\mathcal{D}}(a|s) - \ln_q \pi_{\mathcal{D}}(a|s)\ln \pi_{\mathcal{D}}(a|s)\right] \\ \leq -\pi_{\mathcal{D}}^{q-1}(a|s)\ln \pi_{\mathcal{D}}(a|s) \leq -\pi_{\mathcal{D}}^{q-1}(a|s)&\left(1 - \frac{1}{\pi_{\mathcal{D}}(a|s)}\right) = \pi_{\mathcal{D}}^{q-2}(a|s) - \pi_{\mathcal{D}}^{q-1}(a|s).\end{aligned} \tag{16}$$

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

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

□

Difference with In-sample Softmax. Our proposal is seemingly similar to in-sample softmax [17]. However, the two methods are radically different: in-sample softmax requires satisfying the dataset support constraint during learning to alleviate the out-of-distribution error, and is motivated by the observation that taking softmax is easier than the hard max operator Eq. (5). On the other hand, we assume the behavior policies are Tsallis policies and as such the actions not present in the dataset are those being truncated. This viewpoint leads to the natural satisfaction of the dataset support constraint.

Furthermore, the assumption naturally leads to the choice of Tsallis regularized backward learning which recovers in-sample softmax as a special case when $q = 1$. Since softmax policies have full support, there always exists non-negligible probabilities for suboptimal actions; while for Eq. (12) the optimal policy 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.

Remark. *Tsallis entropy regularization has not been popular since its proposal in RL [9]. 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 provides a better in-sample algorithm that the constraint is satisfied implicitly.*

6 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. In practice, computing the thresholding function ψ requires sorting in the discrete control setting and it remains unclear how to accurately estimate it for the continuous control problems. We found it improves the performance by simply removing ψ and renormalize the policy. In this case, thresholding can still be achieved by simply learning $Q_\theta \to 0$. That is, the learned policy form becomes $\pi_{t+1}(a|s) \propto [Q_\theta(s,a)]_+$. In order to also sample actions from the dataset when updating the policy, we repeat the derivation in Eq. (17) but for the sparsemax policy.

$$\begin{aligned}
\pi_{\pi_{\mathcal{D}},k+1}(a|s) &= \pi_{\mathcal{D}}(a|s)\pi_{\mathcal{D}}(a|s)^{-1} \exp_q Q_\theta(s,a) \\
&= \pi_{\mathcal{D}}(a|s) \exp_q \left(\ln_q \frac{1}{\pi_{\mathcal{D}}(a|s)} \right) \exp_q Q_\theta(s,a) \\
&= \pi_{\mathcal{D}}(a|s)^{\ q-1} \sqrt{\left(\exp_q \left(Q_\theta(s,a) + \ln_2 \frac{1}{\pi_{\mathcal{D}}(a|s)} \right)^{q-1} - (q-1)^2 Q_\theta(s,a) \ln_q \frac{1}{\pi_{\mathcal{D}}(a|s)} \right)}.
\end{aligned} \tag{18}$$

In the last step we made use of the relationship $\left(\exp_q x \cdot \exp_q y\right)^{q-1} = \exp_q(x+y)^{q-1} + (q-1)^2 xy$ [18]. Similar to the discussion after Eq. (9), 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, which gives the following Tsallis in-sample sparsemax actor-critic update rule:

$$\mathcal{L}_{\text{actor}}(\phi) = -\mathbb{E}_{s,a\sim\mathcal{D}} \left[\left(\exp_q \left(Q_\theta(s,a) + \ln_2 \frac{1}{\pi_\omega(a|s)} \right) - Q_\theta(s,a) \ln_q \frac{1}{\pi_\omega(a|s)} \right) \ln \pi_\phi(a|s) \right], \tag{19}$$

$$\mathcal{L}_{\text{baseline}}(\zeta) = \mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\phi(s)} \left[\left(v_\zeta(s) - Q_\theta(s,a) - \tau \ln_2 \pi_\phi(a|s) \right)^2 \right], \tag{20}$$

$$\mathcal{L}_{\text{critic}}(\theta) = \mathbb{E}_{s,a,r,s'\sim\mathcal{D}} \left[\left(r + \gamma v_\zeta(s') - Q_\theta(s,a) \right)^2 \right]. \tag{21}$$

References

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

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

[3] A. Chan, H. Silva, S. Lim, T. Kozuno, A. R. Mahmood, and M. White. Greedification operators for policy optimization: Investigating forward and reverse kl divergences. *Journal of Machine Learning Research*, 23(253):1–79, 2022.

[4] 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.

[5] M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized Markov decission processes. In *36th International Conference on Machine Learning*, volume 97, pages 2160–2169, 2019.

[6] J. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Grundlehren Text Editions. Springer Berlin Heidelberg, 2004.

[7] T. Kozuno, E. Uchibe, and K. Doya. Theoretical analysis of efficiency and robustness of softmax and gap-increasing operators in reinforcement learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 2995–3003, 2019.

[8] T. Kozuno, W. Yang, N. Vieillard, T. Kitamura, Y. Tang, J. Mei, P. Ménard, M. G. Azar, M. Valko, R. Munos, O. Pietquin, M. Geist, and C. Szepesvári. Kl-entropy-regularized rl with a generative model is minimax optimal, 2022. URL `https://arxiv.org/abs/2205.14211`.

[9] 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.

[10] 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.

[11] 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.

[12] O. Nachum and B. Dai. Reinforcement learning via fenchel-rockafellar duality. 2020. URL `http://arxiv.org/abs/2001.01866`.

[13] 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`.

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

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

[16] 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.

[17] 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`.

[18] 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.

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