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.

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, 9, 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 [7, 13]. Convergence properties of the resulting regularized algorithms have been well studied [8, 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). We can define the Tsallis entropy $S_q(\pi)$ in a similar way to Shannon entropy with q-logarithm :

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

When $q = 2$, we recover the Tsallis sparse entropy [12], which achieves the sparsity by projecting onto the probability simplex [2]. In RL, it has been investigated by [10, 11] and it is known that the optimal regularized policy takes the form:

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

$$\psi\left(\frac{Q_\pi(s,\cdot)}{\tau}\right) \doteq \frac{\sum_{a \in K(s)} \frac{Q(s,a)}{\tau} - 1}{|K(s)|},$$

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.

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. (going to put some figures here to show q-statistics behavior and illustrate how the sparsemax policies truncate actions.)

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:

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

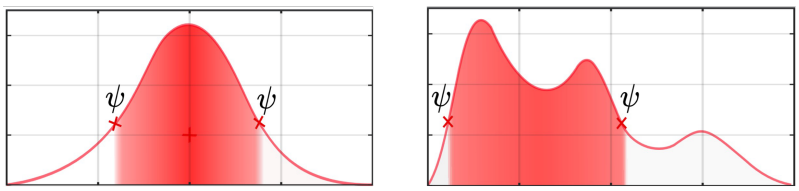

Figure 1: The sparsemax operator acting upon Gaussian and Boltzmann policies by truncating actions with values lower than ψ. The truncated actions can be seen as not present in the offline dataset.

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 Tsallis Regularized Policy for Offline RL

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 actions not present in the dataset are with low probability and are exactly these actions that would have been truncated if by an application of the sparsemax operator.

Let us recall the definition of the sparsemax policy takes the form $\pi_{k+1}(a|s) = \left[\frac{Q_k(s,a)}{\tau} - \psi\left(\frac{Q_k(s,\cdot)}{\tau}\right)\right]_+$. That is, actions with values lower than ψ are set to zero probability; actions with values greater than ψ are included in the set $K(s)$. If we follow the assumption that the

actions present in the dataset are these actions in $K(s)$, we can replace the dataset support constraint $a : \pi_{\mathcal{D}}(a|s) > 0$ by the constraint $a \in K(s)$.

Let us write out the Bellman optimality equation similar to Eq. (6), but replacing $\mathcal{H}(\pi)$ to the Tsallis sparse entropy $S_2(\pi)$:

$$
\begin{aligned}
Q_{*,\pi_{\mathcal{D}}} &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\sum_{a'} \pi(a'|s') \left(Q_{*,\pi_{\mathcal{D}}}(s',a') - \tau S_2(\pi)(\cdot|s') \right) \right] . \\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\frac{1}{2} \tau \left(\sum_{a' \in K(s)} \left(\frac{Q_{*,\pi_{\mathcal{D}}}(s',a')}{\tau} \right)^2 - \psi \left(\frac{Q_{*,\pi_{\mathcal{D}}}(s',\cdot)}{\tau} \right)^2 + 1 \right) \right] \\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\frac{1}{2} \tau \left(\sum_{a' : \pi_{\mathcal{D}}(a'|s') > 0} \left(\frac{Q_{*,\pi_{\mathcal{D}}}(s',a')}{\tau} \right)^2 - \psi \left(\frac{Q_{*,\pi_{\mathcal{D}}}(s',\cdot)}{\tau} \right)^2 + 1 \right) \right] .
\end{aligned}
\tag{10}
$$

Here, $\max_{\pi \in \Delta(\mathcal{A})} \pi(a|s) Q(s,a) + \tau S_2(\pi(\cdot|s))$ attains its maximum at $V(s) = \frac{\tau}{2} \sum_{a \in K(s)} \left(\frac{Q(s,a)}{\tau} \right)^2 - \psi \left(\frac{Q(s,\cdot)}{\tau} \right)^2 + \frac{\tau}{2}$ [10, 12]. Since we assume the conditions $a : \pi_{\mathcal{D}}(a|s) > 0$ by the constraint $a \in K(s)$ are interchangeable, the dataset support constraint is implicitly satisfied by the Tsallis sparse entropy regularization. (we could add something here to say actually for all $q \geq 2$ the policies are variants of sparsemax, but not sure this is needed or not.)

Remark. *Tsallis entropy regularization has not been popular since its proposal in RL [10]. 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.*

5.1 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

$$
\pi_{\pi_{\mathcal{D}},Q_\theta}(a|s) = \pi_{\mathcal{D}}(a|s) \exp \left(\frac{Q_\theta(s,a) - Z(s)}{\tau} - \ln \pi_\omega(a|s) \right) , \tag{11}
$$

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 for $q = 2$. In order to also sample actions from the dataset when updating the policy, we repeat the derivation in Eq. (11) but for the sparsemax policy. Let us define $z_{\pi_{\mathcal{D}},k,\tau}(s,a) := \frac{Q_{\pi_{\mathcal{D}},k}(s,a)}{\tau} - \psi \left(\frac{Q_{\pi_{\mathcal{D}},k}(s,\cdot)}{\tau} \right)$ for convenience:

$$
\begin{aligned}
\pi_{\pi_{\mathcal{D}},k+1}(a|s) &= \pi_{\mathcal{D}}(a|s) \pi_{\mathcal{D}}(a|s)^{-1} \exp_2 z_{\pi_{\mathcal{D}},k,\tau}(s,a) \\
&= \pi_{\mathcal{D}}(a|s) \exp_2 \left(\ln_2 \frac{1}{\pi_{\mathcal{D}}(a|s)} \right) \exp_2 z_{\pi_{\mathcal{D}},k,\tau}(s,a) \\
&= \pi_{\mathcal{D}}(a|s) \left(\exp_2 \left(z_{\pi_{\mathcal{D}},k,\tau}(s,a) + \ln_2 \frac{1}{\pi_{\mathcal{D}}(a|s)} \right) - z_{\pi_{\mathcal{D}},k,\tau}(s,a) \ln_2 \frac{1}{\pi_{\mathcal{D}}(a|s)} \right) .
\end{aligned}
\tag{12}
$$

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].

In the discrete case, the truncation threshold or the normalization function ψ is computed by sorting the action values. However, in the continuous case we can no longer do that. Instead, we propose to use a single-point estimate:

$$V(s) \approx \frac{1}{2}\tau\left(\left(\frac{Q_{\pi_\mathcal{D},k}(s',a')}{\tau}\right)^2 - \psi\left(\frac{Q_{\pi_\mathcal{D},k}(s',\cdot)}{\tau}\right)^2 + 1\right)$$
$$\Leftrightarrow \psi\left(\frac{Q_{\pi_\mathcal{D},k}(s,\cdot)}{\tau}\right) \approx \sqrt{\left(\frac{Q_{\pi_\mathcal{D},k}(s',a')}{\tau}\right)^2 - \frac{V(s)-\frac{1}{2}\tau}{\frac{1}{2}\tau}}.$$

(need to record this ψ value in the experiment to see how big area of the action Gaussian distribution is truncated.) The single-point estimate is similar to the common practice of entropy-regularized actor-critic algorithms such as soft actor-critic [6] that approximates the value function by a single-point estimate, i.e. $V_\zeta(s) \approx Q_\theta(s,a) + \ln \pi_\phi(a|s)$.

We now summarize the loss functions for implementing the Tsallis in-sample actor-critic algorithm:

$$\psi\left(\frac{Q_{\pi_\mathcal{D},\theta}(s,\cdot)}{\tau}\right) \leftarrow \sqrt{\left(\frac{Q_\theta(s,a)}{\tau}\right)^2 - \frac{V_\zeta(s)-\frac{\tau}{2}}{\frac{\tau}{2}}}, \tag{13}$$

$$z_{\pi_\mathcal{D},\theta,\tau}(s,a) \leftarrow \frac{Q_\theta(s,a)}{\tau} - \psi\left(\frac{Q_\theta(s,\cdot)}{\tau}\right), \tag{14}$$

$$\mathcal{L}_{\text{actor}}(\phi) = -\mathbb{E}_{s,a\sim\mathcal{D}}\left[\left(\exp_2\left(z_{\pi_\mathcal{D},\theta,\tau}(s,a) + \ln_2\frac{1}{\pi_\omega(a|s)}\right) - z_{\pi_\mathcal{D},\theta,\tau}(s,a)\ln_q\frac{1}{\pi_\omega(a|s)}\right)\ln\pi_\phi(a|s)\right], \tag{15}$$

$$\mathcal{L}_{\text{baseline}}(\zeta) = \mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\phi(s)}\left[\left(v_\zeta(s) - Q_\theta(s,a) - \tau\ln_2\pi_\phi(a|s)\right)^2\right], \tag{16}$$

$$\mathcal{L}_{\text{critic}}(\theta) = \mathbb{E}_{s,a,r,s'\sim\mathcal{D}}\left[\left(r + \gamma v_\zeta(s') - Q_\theta(s,a)\right)^2\right]. \tag{17}$$

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] T. Haarnoja, A. Zhou, P. Abbeel, and S. 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*, pages 1861–1870, 2018.

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

[8] 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.

[9] 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`.

[10] 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.

[11] 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.

[12] 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.

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

[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.