2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Generalized Maximum Entropy Differential Dynamic Programming

Yuichiro Aoyama[1,2] and Evangelos A. Theodorou[1]

***Abstract*— We present a sampling-based trajectory optimization method derived from the maximum entropy formulation of Differential Dynamic Programming with Tsallis entropy. This method is a generalization of the legacy work with Shannon entropy, which leads to a Gaussian optimal control policy for exploration during optimization. With the Tsallis entropy, the policy takes the form of q-Gaussian, which further encourages exploration with its heavy-tailed shape. Moreover, the sampling variance is scaled according to the value function of the trajectory. This scaling mechanism is the unique property of the algorithm with Tsallis entropy in contrast to the original formulation with Shannon entropy, which scales variance with a fixed temperature parameter. Due to this property, our proposed algorithms can promote exploration when necessary, that is, the cost of the trajectory is high. The simulation results with two robotic systems with multimodal cost demonstrate the properties of the proposed algorithm.**

I. Introduction

Tsallis entropy, also known as q-logarithmic entropy, is a generalization of the standard Boltzmann-Gibbs or Shannon entropy [1], [2]. The Tsallis entropy is nonadditive, which means that the sum of the entropy of probabilistically independent subsystems is not equal to that of the entire system. The entropy is used in nonextensive statistical mechanics which can handle strongly correlated random variables [3], [4]. It also helps to analyze complex phenomena in physics such as in [5], [6], etc.

Maximization of Shannon entropy under fixed mean and covariance yields a Gaussian distribution [7]. With the Tsallis entropy, similar maximization with a fixed q-mean and covariance leads to a q-Gaussian distribution [8], [9]. The distribution has a heavy tail compared to a normal Gaussian with a certain range of the entropic index q. Due to this property, the q-Gaussian distribution is utilized for several engineering applications. In [10], a mixture of the distribution is used for image and video semantic mapping, improving the robustness to outliers. In stochastic optimization, the distribution is used as a generalization of the Gaussian kernel, showing better control ability in smoothing functions [11]. It was also used for mutation in the evolutionary algorithm, showing its effectiveness over other distributions [12].

Maximum Entropy (ME) is a popular technique in various fields, such as Stochastic Optimal Control (SOC) and Reinforcement Learning (RL), which can improve the robustness of stochastic policies. In ME, an entropic regularization term is added to the objective, encouraging exploration during optimization and preventing policies from converging to a delta distribution. [13], [14]. This is because the regularized objective simultaneously maximizes entropy while minimizing the original objective. In SOC, Shannon entropy, which leads to Kullback-Leibler (KL) divergence, is used as a regularization term between the controlled and prior distributions [15]. In [16], Tsallis divergence, a generalization of the KL divergence, is used, showing improvements in robustness. For RL application, Tsallis entropic regularization leads to better performance and faster convergence [17]–[19]. Recently, the ME technique with Shannon entropy has been applied to a trajectory optimization algorithm Differential Dynamic Programming (DDP) [20], which yields a new algorithm ME-DDP [21]. DDP is a powerful trajectory optimization tool that has a quadratic convergence rate [22]. However, since it relies on local information, of cost and dynamics, it converges to local minima rather than global minima. In contrast, ME-DDP can explore multiple local minima while minimizing the original objective. Consequently, it can find better local minima than those of normal DDP.

In this paper, we propose ME-DDP with Tsallis entropy, which is a generalization of ME-DDP with Shannon entropy. The Tsallis entropy turns the optimal policy from Gaussian to heavy-tailed q-Gaussian, improving the exploration capacity of ME-DDP. Moreover, the generalized formulation can scale the variance of the policy based on the value function of the trajectory, which further promotes exploration. We validate our proposed algorithm in two robotic systems, i.e., a 2D car and a quadrotor, and make comparison with normal DDP and ME-DDP with Shannon entropy.

The main contribution of this paper is as follows.

- We derive DDP with Tsallis entropic regularization.
- We show the superior exploration capability of ME-DDP with Tsallis entropy to ME-DDP with Shannon entropy by analyzing the stochastic policies.
- We validate the exploration capability of our method with two robotic systems in simulation.

II. Preliminaries

A. Tsallis Entropy

With an entropic index $q \in \mathbb{R}$, we introduce q-logarithm function

$$\ln_q(x) = \begin{cases} \ln(x), & q = 1,\ x > 0 \\ \frac{x^{1-q}-1}{1-q}, & q \neq 1\ x > 0, \end{cases} \tag{1}$$

and its inverse q-exponential function

$$\exp_q(x) = \begin{cases} \exp(x), & q = 1, \\ [1-(q-1)x]_+^{-\frac{1}{q-1}}, & q \neq 1. \end{cases}$$

[1]School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[2]Komatsu Ltd.,Tokyo, Japan
{yaoyama3, evangelos.theodorou}@gatech.edu

Consider a discrete set of probabilities $\{p_i\}$. $i = 1, \cdots, I$. The Tsallis or q-entropy is defined as

$$S_q(p) = \frac{1 - \sum_i p_i^q}{q-1}, \tag{2}$$

which is a generalization of the Shannon entropy

$$H(p) = \sum_i p_i \log\left(\frac{1}{p_i}\right) = -\sum_i p_i \log p_i.$$

The Tsallis entropy may be represented as a sum of the product of probability and q-log probability as

$$S_q(p) = \sum_i p_i \ln_q\left(\frac{1}{p_i}\right) = \sum_i p_i \frac{p_i^{q-1} - 1}{1-q} = \frac{1 - \sum_i p_i^q}{q-1},$$

which recovers the definition in (2). We believe that this is the most straightforward generalization. Although

$$S_q(p) = -\sum_i p_i \ln_q p_i$$

can be another option for the definition, the following relation

$$-\ln_q p_i \neq \ln_q(1/p_i),$$

changes the parametrization from q to $2-q$ as in [18]. We use the former definition in (2) to avoid this change. The argument above holds with continuous probability distribution by changing the summation to integral.

B. Univariate q-Gaussian distribution

Let us consider a Probability Distribution Function (PDF) $p_q(x)$ of a scalar random variable $x \in \mathbb{R}$. q-Gaussian distribution is a generalized version of Gaussian distribution, which is obtained by maximizing the q-entropy with a fixed (given) q-mean μ_q and q-variance σ_q^2, where the subscript q means that they are computed with the q-escort distribution, which is a normalized q th power of the original $p_q(x)$, i.e.,

$$\mu_q = \frac{\int x p_q(x)^q \mathrm{d}x}{\int p_q(x)^q \mathrm{d}x}, \quad \sigma_q^2 = \frac{\int (x-\mu_q)^2 p_q(x)^q \mathrm{d}x}{\int p_q(x)^q \mathrm{d}x}.$$

This normalization is known to be the correct formulation of nonextensive statistics [9]. The univariate q-Gaussian distribution is given as follows [9].

$$p_q(x) = \frac{1}{Z_q}\left(1 - \frac{1-q}{3-q}\frac{(x-\mu_q)^2}{\sigma_q^2}\right)^{\frac{1}{1-q}},$$

$$\text{with } Z_q = \begin{cases} \left[\sigma_q^2 \frac{3-q}{q-1}\right]^{\frac{1}{2}} B\left(\frac{1}{2}, \frac{3-q}{2(q-1)}\right), & 1 < q < 3, \\ \left[\sigma_q^2 \frac{3-q}{1-q}\right]^{\frac{1}{2}} B\left(\frac{1}{2}, \frac{2-q}{1-q}\right), & q < 1, \end{cases}$$

where $B(\cdot,\cdot)$ is Beta function. The bounds for q are for the convergence of integrals to compute the partition function. Here, we provide two properties of the function.

Compact support when $q < 1$.
To satisfy $p_q(x) \geq 0$, $p_q(x)$ has a compact support in the case of $q < 1$, which is given by $x \in [x_b, x_u]$, with

$$x_b = \mu - a_q, \; x_u = \mu + a_q, \; a_q = \sqrt{(3-q)/(1-q)}.$$

Note that $\mu_q = \mu$ when $q < 2$.
Recovers Normal Gaussian as $q \to 1$.

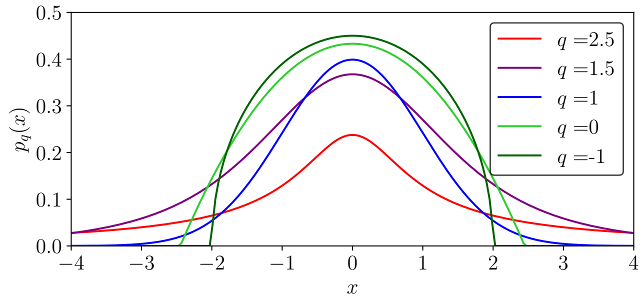


Fig. 1: q-Gaussian distribution with different qs with $\mu_q = 0$ and $\sigma_q^2 = 1$. $q = 1$ corresponds to a normal Gaussian.

Using the definition of exponential $e^x = \lim_{n\to 0}(1 + x/n)^n$, p_q approaches

$$p_1(x) = (1/\sqrt{2\pi}\sigma)\exp[(x-\mu)^2/2\sigma^2],$$

as $q \to 1$, which is normal Gaussian distribution. In addition to these, moments are only defined with certain qs due to the convergence condition of integrals. A sampling method from the distribution is proposed in [23].

We show some $p_q(x)$ with different q in Fig. 1. As can be seen from the figure, a large q with $q > 1$ yields a heavier-tailed distribution than the normal Gaussian ($q = 1$). We can also observe the compact support when $q < 1$. For exploration purposes, we prefer the heavy-tailed distribution and thus focus on the case where $q > 1$, hereafter.

C. Multivariate q-Gaussian Distribution

The multivariate variant of p_q for $q > 1$ and $x \in \mathbb{R}^n$ is

$$p_q(x) = \frac{1}{Z_q}\left[1 + \frac{q-1}{n+2-nq}(x-\mu_q)^\top \Sigma_q^{-1}(x-\mu_q)\right]^{-\frac{1}{q-1}},$$

$$\text{with } Z_q = \left[\frac{(n+2)-nq}{q-1}\right]^{\frac{n}{2}} |\Sigma_q|^{\frac{1}{2}} \pi^{\frac{n}{2}} \frac{\Gamma\left(\frac{1}{q-1} - \frac{n}{2}\right)}{\Gamma\left(\frac{1}{q-1}\right)}, \tag{3}$$

where $\Gamma(\cdot)$ is Gamma function [24], [25], and $\Sigma_q \in \mathbb{R}^{n\times n}$ is q-covariance, which is a multivariate version of σ_q^2. Here, q needs to satisfy the following condition.

$$1 < q < 1 + 2/n. \tag{4}$$

In addition, when the tighter condition

$$1 < q < (n+4)/(n+2) = 1 + 2/(n+2), \tag{5}$$

is satisfied, (normal, not q) covariance $\Sigma \in \mathbb{R}^{n\times n}$ exists and is finite. It has the following relation with Σ_q :

$$\Sigma = \frac{n+2-nq}{n+4-(n+2)q}\Sigma_q,$$

whose coefficient is positive due to (4), (5). Under $1 < q < 1 + 2/(n+1)$, which is slightly looser than (5), $\mu = \mu_q$. Eq. (3) corresponds to the Student's t distribution [7], [26]

$$p_t(x) = \frac{\Gamma(\nu+n)/2}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}|\Sigma_t|}\left[1 + \frac{1}{\nu}(x-\mu_t)^\top \Sigma_t^{-1}(x-\mu_t)\right]^{-\frac{\nu+n}{2}}.$$

ν here is known as degrees of freedom. By taking

$$\nu = \frac{n+2-nq}{q-1}, \quad \frac{q-1}{n+2-nq}\Sigma_q^{-1} = \frac{1}{\nu}\Sigma_t^{-1}, \qquad (6)$$

the q-Gaussian distribution $(q > 1)$ is recovered. We use this property to sample from the distribution in section IV.

D. Maximum Entropy DDP

In this section, we have a brief review of ME-DDP with unimodal and multimodal policies [21]. Consider a trajectory optimization problem of a dynamical system with state $x \in \mathbb{R}^{n_x}$ and control $u \in \mathbb{R}^{n_u}$. Let us define the state and control trajectory with time horizon T as

$$X = [x_0, \cdots, x_T], \quad U = [u_0, \cdots, u_{T-1}],$$

and deterministic dynamics $x_{t+1} = f(x_t, u_t)$. The problem is formulated as a minimization problem of the cost

$$J(X,U) = J(x_0, U) = \sum_{t=0}^{T-1} l_t(x_t, u_t) + \Phi(x_T), \qquad (7)$$

subject to the dynamics. Here, l_t and Φ are running and terminal cost, respectively. We consider a stochastic control policy $\pi(x_t|u_t)$ with the same dynamics. Moreover, we add the Shannon entropic regularization term

$$H[\pi_t] = \mathbb{E}_{\pi_t}[\ln \pi_t] = -\int \pi_t(u_t)[\ln \pi_t]\mathrm{d}u_t,$$

to the cost and consider the expectation of the cost over π

$$J_\pi = J_\pi(x_0, \pi) = \mathbb{E}\big[\Phi(x_T) + \sum_{t=0}^{T-1}\big(l_t(x_t, u_t) - \alpha H[\pi_t]\big)\big],$$

where $\alpha (> 0)$ is a temperature that determines the effect of the regularization term [13]. Applying the Bellman's principle for normal DDP, i.e.,

$$V(x_t) = \min_{u_t}\{l_t(x_t, u_t) + V(x_{t+1})\},$$

with value function V, in our setting, we have

$$V(x_t) = \min_{\pi_t}\big\{\mathbb{E}_\pi[l_t(x_t, u_t) + V(x_{t+1})] - \alpha H[\pi_t]\big\} \qquad (8)$$

The solution to right-hand side of the equation is

$$\min_\pi \mathbb{E}_{u\sim\pi(\cdot|x)}[\underbrace{l(u,x) + V(x')}_{Q(x,u)}] - \alpha H[\pi(\cdot|x)],$$

$$\text{subject to} \int \pi(u|x)\mathrm{d}u = 1, \qquad (9)$$

where we dropped time instance t and denote x_{t+1} as x'. $\pi(\cdot|x)$ means that it originally was a function of u, but it vanishes after taking the expectation to compute H. We emphasize that the expectation is computed by sampling the controls from π by changing the notation in the expectation. The optimization problem above is solved by forming the Lagrangian and setting the functional derivative of π zero. The optimal policy and value function is obtained as

$$\pi^*(u|x) = \frac{1}{Z(x)}\exp\big[-\frac{1}{\alpha}Q(x,u)\big], \qquad (10)$$

$$V(x) = -\alpha \ln Z(x), \qquad (11)$$

with a partition function $Z(x)$.

1) Unimodal policy: To obtain π^*, we first consider a deviation $\delta x, \delta u$ from a nominal trajectory $\bar{x}, \bar{u}$, having a pair $x = \bar{x} + \delta x, u = \bar{u} + \delta u$. Then, we perform a quadratic approximation of Q around $\bar{x}$, $\bar{u}$, and plug it in (10), having a Gaussian policy

$$\pi^*(\delta u|\delta x) \sim \mathcal{N}(\delta u^*, \alpha Q_{uu}^{-1}), \qquad (12)$$

where δu^* is a solution of normal DDP, i.e.,

$$\delta u^* = k + K\delta x, \qquad (13)$$
$$\text{with } k = -Q_{uu}^{-1}Q_u, \ K = -Q_{uu}^{-1}Q_{ux}.$$

As in normal DDP, ME-DDP has backward and forward passes. It can have N (two in the original work) trajectories in parallel. The backward pass is the same as that of the normal DDP. In the forward pass, a new control sequence is sampled from the optimal policy based on the best (lowest cost) trajectory for every m iteration. In the sampling phase, the best trajectory is kept and only the remaining $N-1$ trajectories are sampled. Aside from that, the pass is the same as normal DDP, that is, forward propagation of dynamics and line search for cost reduction.

2) Multimodal policy: Here, we consider LogSumExp approximation of the value function using N trajectories, which gives the terminal state of the value function as

$$V(x_T) = \tilde{\Phi}(x) = -\alpha \ln \sum_{n=1}^{N} \exp\Big[-\frac{1}{\alpha}\Phi^{(n)}(x)\Big],$$

where $\Phi^{(n)}(x)$ $n = 1, \cdots N$, is the terminal cost of nth trajectory. Exponential transformation $\mathcal{E}_\alpha(y) = \exp(-y/\alpha)$, of the value function allows us to write Z as a sum of those of N trajectories denoted by $z^{(n)}$,

$$z(x) = \mathcal{E}_\alpha[V(x)] = \exp\big(-(-\alpha \ln Z(x))/\alpha\big)$$
$$= Z(x) = \sum_{n=1}^{N} z^{(n)}(x).$$

Let us also transform the running cost, having the desirability function $r(x,u) = \mathcal{E}_\alpha[l(x,u)]$. Due to the property of the exponential, i.e., $\exp(x+y) = \exp(x)\exp(y)$, the optimal policy becomes

$$\pi(u|x) = \frac{1}{Z}\exp\Big[-\frac{1}{\alpha}(\underbrace{l(u,x) + V(x')}_{Q(u,x)})\Big] = \frac{1}{Z}z(x')r(x,u).$$

Thanks to these properties, we obtain the optimal policy as a weighted sum of those of each trajectory $\pi^{(n)}$s as

$$\pi(u|x) = \sum_{n=1}^{N} w^{(n)}(x)\pi^{(n)}(u|x),$$

with weight $w^{(n)} \propto \mathcal{E}_\alpha[V^{(n)}(x)]$. Since $\pi^{(n)}$ are Gaussian as in (12), the policy is now a mixture of Gaussians (and thus multimodal) with weights based on value functions.

III. Generalized Max Entropy DDP

A. Tsallis entropic regularization

Based on ME-DDP in the previous section, we now consider entropic regularization with Tsallis entropy

$$S_q = \frac{1 - \int p^q(x)\mathrm{d}x}{q-1}.$$

Let us revisit the optimization problem in (9). We now use S_q instead of H as a regularization term, and consider the q-escort distribution of π with a normalization constant C, which transforms the problem as

$$\min_{\pi} \mathbb{E}_{\pi^q(\cdot|x)/C}[l(u,x) + V(x')] - \alpha S_q[\pi(\cdot|x)], \tag{14}$$

under the same constraints for a valid PDF for π^q/C as in (9). To solve this optimization problem, we form a Lagrangian as

$$\mathcal{L} = \int Q\frac{\pi^q}{C}\mathrm{d}u - \alpha S_q[\pi] + \lambda\Big(1 - \int \pi \mathrm{d}\pi\Big),$$

with a Lagrangian multiplier λ. The first-order optimality condition, i.e., $\nabla_\pi \mathcal{L} = 0$ gives the optimal policy as

$$\pi^* = Z^{-1}\Big[(q-1)\frac{Q}{C\alpha} + 1\Big]^{-\frac{1}{q-1}}, \tag{15}$$

where Z is a partition function. By plugging this back into (14) and using (8), the value function is obtained as follows.

$$\begin{aligned}
&V(x) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (16)\\
=&Z^{-q}\int Q\frac{\pi^{*q}}{C}\mathrm{d}u - \alpha\frac{1-\int \pi^{*q}\mathrm{d}u}{q-1}.\\
=&Z^{-q+1}\frac{\alpha}{q-1}\int \underbrace{Z^{-1}\Big[(q-1)\frac{Q}{C\alpha}+1\Big]^{-\frac{1}{q-1}}}_{\pi^*}\mathrm{d}u - \frac{\alpha}{q-1}\\
=&-\alpha\frac{Z^{1-q}-1}{1-q} = -\alpha \ln_q Z \quad (\because (1)).
\end{aligned}$$

Notice that this expression is obtained by changing $\log$ in (11) to q-log, which implies that the derivation is the generalization of ME-DDP with Shannon to Tsallis entropy.

B. q-Gaussian Policy

To examine the property of π^*, We perform quadratic approximation of Q around nominal trajectories $(\bar{x}, \bar{u})$ and complete the square as

$$Q(x,u) = Q(\bar{x},\bar{u}) + \delta Q(\delta x, \delta u) = \tilde{V}(x) + \frac{1}{2}v^\mathsf{T}Q_{uu}v,$$
$$\text{with} \quad \tilde{V}(x) = Q(\bar{x},\bar{u}) + \delta Q(\delta x, \delta u^*),\ v = \delta u - \delta u^*.$$

$\tilde{V}(x)$ is the value function of normal DDP, which is obtained by plugging δu^* into $Q(x,u)$. δQ contains terms up to the second order. Substituting these back into (15), we have

$$\begin{aligned}
\pi^* &= Z^{-1}\Big[\frac{q-1}{C\alpha}\Big(\tilde{V}(x) + \frac{1}{2}v^\mathsf{T}Q_{uu}v\Big) + 1\Big]^{-\frac{1}{q-1}} \qquad (17)\\
&= Z^{-1}\Big[\frac{(q-1)\tilde{V}(x) + C\alpha}{C\alpha}\Big]^{-\frac{1}{q-1}}\\
&\times\Big[1 + \frac{q-1}{n_u+2-n_uq}v^\mathsf{T}\frac{(n_u+2-n_uq)}{2[(q-1)\tilde{V}(x)+C\alpha]}Q_{uu}v\Big]^{-\frac{1}{q-1}},
\end{aligned}$$

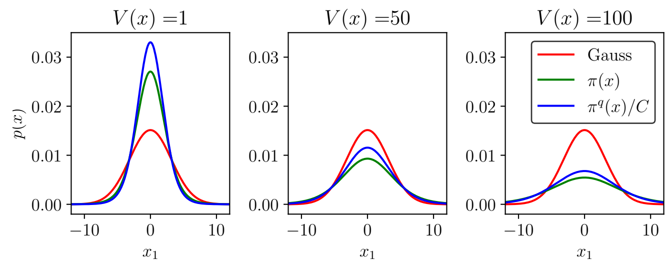


Fig. 2: Normal Gaussian policy, q-Gaussian policy π and q-escort distribution of π^q/C policy with different value function.

which is a q-Gaussian (see (3) and change of n to n_u) with

$$\mu_q = \delta u^*, \quad \Sigma_q = \frac{2[(q-1)\tilde{V}(x) + C\alpha]}{(n_u+2-n_uq)}Q_{uu}^{-1}. \tag{18}$$

Here, C, the normalization term for the escort distribution, has not been determined and does not have a closed-form solution. C is obtained by solving the following equation.

$$\begin{aligned}
\int \pi^q\mathrm{d}u = C \Leftrightarrow \Big[\tilde{V}(x) + \frac{\alpha C}{q-1}\Big]^{\frac{n_u}{2}(q-1)} C \qquad (19)\\
= \frac{n_u+2-n_uq}{2}\Bigg[|Q_{uu}^{-1}|^{\frac{1}{2}}(2\pi)^{\frac{n_u}{2}}\frac{\Gamma\big(\frac{1}{q-1} - \frac{n_u}{2}\big)}{\Gamma\big(\frac{1}{q-1}\big)}\Bigg]^{(1-q)}.
\end{aligned}$$

Since the left-hand side is monotonically increasing with C, and since C is a scalar, the equation can be easily solved by numerical methods, such as bisection search. Observe in (18) that we can recover ME-DDP with Shannon entropy by $q \to 1$, which leads to π^q/C (q-Gaussian) $\to \pi$ (Gaussian), and $\Sigma_q \to \alpha Q_{uu}^{-1}$.

C. Sampling from q-escort distribution

In the forward pass of ME-DDP with Shannon entropy, a new control sequence is sampled from π which is a Gaussian or a Gaussian mixture. In our case, since expectation of the cost is taken over the q-escort distribution of q-Gaussian π, it is more natural to sample from π^q/C than from π. To sample from this distribution, we use the property of q-Gaussian, that is, the q-escort distribution of a q-Gaussian is also a q-Gaussian. To see this, let us introduce a parameter q' and a q-mean and covariance as follows.

$$q' = 2 - \frac{1}{q},\ \mu'_q = \mu_q,\ \Sigma'_q = \frac{n_u+2-n_uq}{n_u+(2-n_u)q}\Sigma_q. \tag{20}$$

With these, consider a PDF parameterized by q' as

$$p_{q'}(x) = \frac{1}{Z_{q'}}\Big[1 - \frac{1-q'}{n_u+2-n_uq'}y^\mathsf{T}[\Sigma'_q]^{-1}y\Big]^{\frac{1}{1-q'}},$$

with $y = x - \mu_{q'}$. By substituting (20) in, we see that $p_{q'}(x)$ is proportional to

$$\Big(\Big[1 - \frac{1-q}{n_u+2-n_uq}(x-\mu_q)^\mathsf{T}\Sigma_q^{-1}(x-\mu_q)\Big]^{\frac{1}{1-q}}\Big)^q,$$

which is a q th power of q-Gaussian. This implies that sampling from the q-escort distribution $(p_q(x))^q/C$ is achieved by sampling from the q-Gaussian distribution $p_{q'}(x)$ obtained by the transformation given in (20). Moreover, sampling from q-Gaussian is equivalent to sampling from Student's t distribution with the transformation in (6) [27]. We use the technique to sample from the π^q/C in our algorithm.

Here, we analyze the difference between the optimal control policies of ME-DDP with Shannon and Tsallis entropy. With Shannon entropy, the covariance of Gaussian π^* is determined by αQ_{uu}^{-1} as in (12). On the other hand, with Tsallis entropy, the covariance is also affected by the value function as in (18). We can interpret the existence of the value function as follows. When the cost of the trajectory is low and thus $\tilde{V}(x)$ is low, the algorithm does not need to explore much because the current trajectory is already good. Therefore, the covariance for exploration is small. In the opposite case, the covariance is amplified by the large $\tilde{V}(x)$, which encourages exploration. We visualize the normal Gaussian policy, q-Gaussian policy π, and its q-escort distribution π^q/C with a 2D state $x = [x_1, x_2]^\mathsf{T} \in \mathbb{R}^2, \alpha = 10$ and a unit Q_{uu} in Fig. 2 for better understanding. As shown in the top of the figure, the three panels correspond to three different $\tilde{V}(x)$ s. On the left, when $\tilde{V}(x)$ is small, π^q is even tighter than Gaussian. While the Gaussian policy remains the same shape, π^q becomes heavy-tailed as $\tilde{V}(x)$ increases to the right. This means that the covarinace is properly scaled based on the cost of the trajectory, rather than scaled with the same scaling factor. Due to this property, we deduce that the ME-DDP with Tsallis entropy case has better exploration capability, which we validate in section IV.

The proposed algorithm is summarized in Alg. 1. It takes in N nominal control sequences and performs optimization, outputting one pair of state control trajectories that achieve the lowest cost. In the algorithm, we write Q_{uu}^{-1} as Σ so that it can be easily compared with ME-DDP with Shannon entropy.

D. Availability of multimodal policy

In ME-DDP with Shannon entropy, the multimodal policy is available due to the additive structure of the partition function explained in section II-D.2. However, with Tsallis entropy, this is not the case. Indeed, from the partition function in (17), we obtain

$$
\begin{aligned}
z &= \int \Big[(q-1)\frac{Q}{C\alpha}+1\Big]^{-\frac{1}{q-1}}\mathrm{d}u \\
&= \int \Big[1-(q-1)\Big(\frac{l(x,u)+V(x')}{C\alpha}\Big)\Big]^{-\frac{1}{q-1}}\mathrm{d}u \\
&= \int \exp_q\Big(\frac{-l(x,u)}{C\alpha}\Big)\otimes_q \exp_q\Big(\frac{-V(f(x,u))}{C\alpha}\Big)\mathrm{d}u \\
&= \int r(x,u)\otimes_q z(f(x,u))\mathrm{d}u,
\end{aligned}
$$

where $\otimes_q$ is q-product [25] that is not distributed

$$(a+b)\otimes_q c \neq a\otimes_q c + b\otimes_q c.$$

Therefore, z is not written as a sum of $z^{(n)}$s, as opposed to the Shannon entropy case.

Algorithm 1: Generalized ME-DDP

Input: x_0 : initial state, $u^{(1:N)}$, $\Sigma^{(1:N)}$, $K^{(1:N)}$: initial sequence, m : sampling frequency, I: max iteration, q: entropic index that satisfies (4) .
Result: $x^{(b)}, u^{(b)}, K^{(b)}$
Compute initial trajectory $(x_{0:T}^{(1:N)})$ and cost $J^{(1:N)}$.
for $i \leftarrow 0$ *to* I **do**
  if $i\%m = 0$ **then**
    $x^{(1)}, u^{(1)}, K^{(1)}, \Sigma^{(1)}, J^{(1)} \leftarrow$ lowest cost mode
    $\pi \leftarrow q$-Gaussian($u^{(1)}$, $x^{(1)}$, $K^{(1)}$, $\Sigma^{(1)}$);
    for $n \leftarrow 2$ *to* N *in parallel* **do**
      $x^{(n)}, u^{(n)}, K^{(n)}, J^{(n)}$ $\leftarrow$ Sample from π^q/C with (19), (20) and sampling technique from Student's t distribution in [27].
    end
  end
  for $n \leftarrow 1$ *to* N *in parallel* **do**
    $k^{(n)}, K^{(n)}, \Sigma^{(n)} \leftarrow$ Backward Pass
    $x^{(n)}, u^{(n)}, J^{(n)} \leftarrow$ Line Search.
  end
end
$b \leftarrow \arg\min_n J^{(n)}$

IV. Numerical Experiments

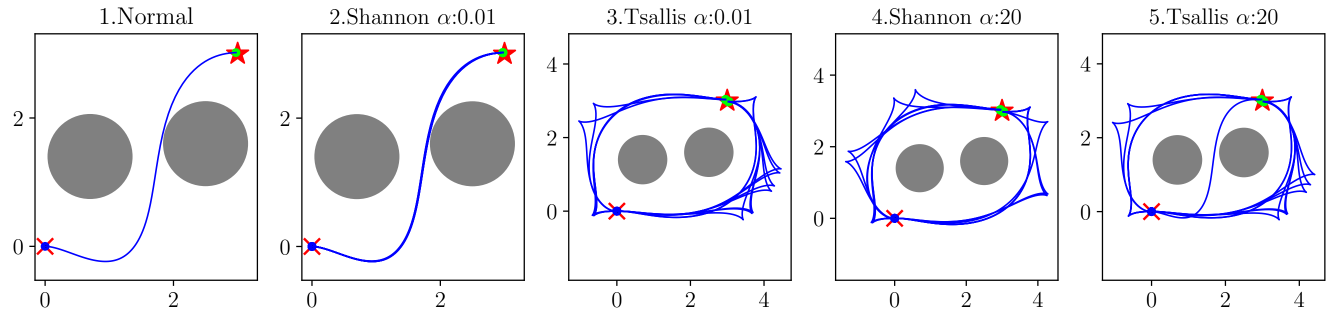
In this section, we validate our proposed algorithm using two systems, a 2D car and a quadrotor. The tasks are to reach the targets while keeping distance from spherical obstacles which are encoded as part of the cost in (7). We first give a belief description of the cost structure and systems. Then, we provide results of the experiment, including comparison with a normal DDP without entropic regularization, ME-DDP with Shannon entropy, and our proposed ME-DDP with Tsallis entropy. In the experiment, we also examine the effect of the temperature α. Although ME-DDP with Shannon entropy has unimodal and multimodal versions, we use the multimodal one for comparison because of its superior performance [21]. We note that ME-DDP with Tsallis entropy has a unimodal policy, as explained in section III-D. In both algorithms, the number of trajectories is $N = 8$.

A. Cost structure

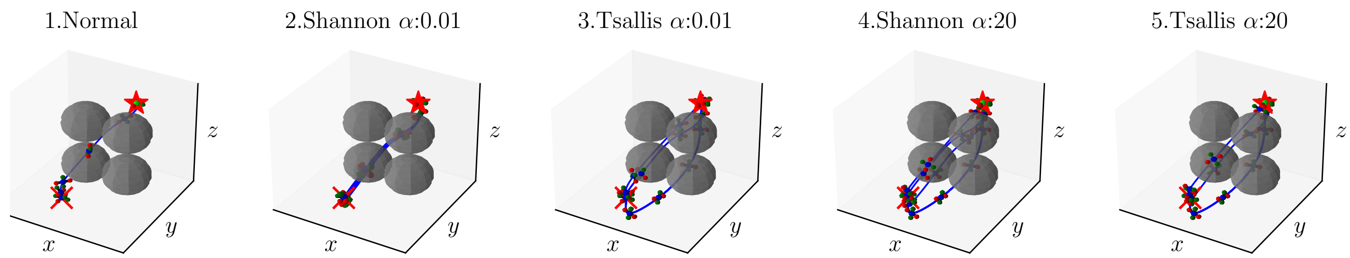
The cost for obstacles is given by

$$l_{\mathrm{s}}(x_t) = \exp\big(-(x_t - c_{\mathrm{o}})^2/2r_{\mathrm{o}}^2\big), \tag{21}$$

where c_{o} and r_{o} are the center and radius of an obstacle, respectively. This cost structure is used to validate ME-DDP with Shannon entropy [21]. We use quadratic cost for the running and terminal costs in (7) and add them with (21). The obstacle term introduces several Gaussian-shaped hills to the landscape of the cost, introducing multiple local minima.

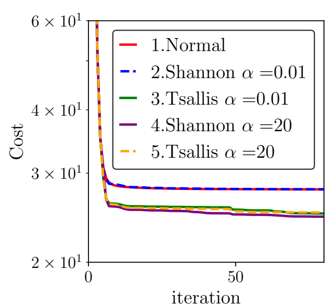


(a) 2D Car Example.

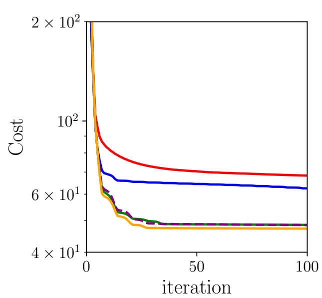


(b) Quadrotor Example.

Fig. 3: Comparison of normal DDP, multimodal ME-DDP with Shannon entropy, and ME-DDP with Tsallis entropy with two different αs. Trajectories of the 15 experiments are overlaid. × and ★ indicate the initial and target positions, respectively. The obstacles are drawn in gray. In the quadrotor example, obstacles are drawn in transparent.



(a) 2D Car



(b) Quadrotor

Fig. 4: Evolution of mean cost over iterations. The numbers in the legends corresponds to those in Fig.3

B. 2D car

The state consists of the position and orientation θ of the car, and thus the state $x = [p_x, p_y, \theta] \in \mathbb{R}^3$. The control u is translation and angular velocities $u = [v, \omega] \in \mathbb{R}^2$. In ME-DDP with Tsallis entropy, q must satisfy $1 < q < 2$ from (4). We choose $q = 1.8$. The control is initialized with all zeros and thus the car stays at a initial position.

C. Quadrotor

The state of the system consists of position, velocity, orientation, and angular velocity, all of which are $\mathbb{R}^3$, and thus $x \in \mathbb{R}^{12}$. The control u of the system is the force generated by the four rotors, which gives $u \in \mathbb{R}^4$. The dynamics is found in [28]. The requirement for q is $1 < q < 1.5$, and we choose $q = 1.4$. The control sequence is initialized with hovering. This sequence is obtained by setting each rotor force $mg_0/4$ where m and g_0 are the mass of the quadrotor and gravitational acceleration.

D. Results

Fig. 3 shows the trajectories of the experiments with normal DDP, multimodal ME-DDP with Shannon entropy, and ME-DDP with Tsallis entropy with two different temperature αs. The experiments of ME-DDPs are performed 15 times, generating 15 trajectories. Fig. 4 shows the evolution of the mean cost of experiments over optimization iterations.

In both dynamics, the normal DDP finds a local minimum that goes through between obstacles. Although it can let trajectories hit targets, the corresponding costs are high because trajectories get close to the obstacles. However, ME-DDPs can explore and find better local minima that keep a greater distance from obstacles. ME-DDP with Shanon entropy requires a large α to explore (see 2s and 4s in Fig. 3). In fact, with a small α, the results are almost the same as those of normal DDP (see 1s and 2s in Fig. 3). This is because α determines the extent of exploration by scaling the sampling variance of the policy. With generalized entropy, ME-DDP can explore well even with small αs (see 3s and 5s in Fig. 3). Although exploration is carried out with a unimodal policy in ME-DDP with Tsallis entropy, its exploration capability is better or equivalent to the multimodal policy of ME-DDP with Shannon entropy. This is the effect of the heavy-tailed shape of q-Gaussian and the value function on variance (18) analyzed in section III. Even with a small α, the value function amplifies the variance for exploration when the value function (and thus the cost) of the trajectory is high, encouraging exploration. Due to this property, tuning α of generalized ME-DDP is easy because one can choose a small α, rather than finding a good α.

With a large $\alpha(=20)$, ME-DDP with Tsallis entropy fails to find a good local minimum, which it used to find (see 3 and 5 in Fig. 3a). We speculate that this is because the variance is scaled too much, making it difficult to sample meaningful trajectories in the exploration process. We have observed that the ME-DDP with Shanon entropy has the same tendency with too large α during the experiment.

V. Conclusion

In this paper, we derived ME-DDP with Tsallis entropy. The algorithm has q-Gaussian policy whose sampling variance is scaled not only by the fixed temperature parameter α, but also by the value function of the trajectory that is optimized. We tested our algorithm using two robotic trajectory optimization problems whose objectives have multiple local minima. The results show that although our algorithm can only have a unimodal policy, it can find better local minima even with a small temperature parameter α compared to the multimodal ME-DDP with Shannon entropy.

Future work includes hardware implementation of the algorithms, as well as theoretical proofs such as the rate and condition of convergence. We are also interested in deriving ME-DDP with different entropy such as Rényi entropy.

Acknowledgement

This work is supported by NASA under the ULI grant 80NSSC22M0070.

References

[1] C. Tsallis, "Possible generalization of boltzmann-gibbs statistics," *Journal of Statistical Physics*, vol. 52, no. 1, pp. 479–487, Jul 1988. [Online]. Available: https://doi.org/10.1007/BF01016429

[2] T. Leinster, *Entropy and Diversity: The Axiomatic Approach*. Cambridge University Press, 2021.

[3] C. Tsallis, M. Gell-Mann, and Y. Sato, "Asymptotically scale-invariant occupancy of phase space makes the entropy sq extensive," *Proc Natl Acad Sci U S A*, vol. 102, no. 43, pp. 15 377–15 382, Oct. 2005.

[4] S. Umarov, C. Tsallis, and S. Steinberg, "On a q-central limit theorem consistent with nonextensive statistical mechanics," *Milan Journal of Mathematics*, vol. 76, no. 1, pp. 307–328, Dec 2008. [Online]. Available: https://doi.org/10.1007/s00032-008-0087-y

[5] E. Lutz, "Anomalous diffusion and tsallis statistics in an optical lattice," *Phys. Rev. A*, vol. 67, p. 051402, May 2003. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.67.051402

[6] B. Liu and J. Goree, "Superdiffusion and non-gaussian statistics in a driven-dissipative 2d dusty plasma," *Phys. Rev. Lett.*, vol. 100, p. 055003, Feb 2008. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.100.055003

[7] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, January 2006. [Online]. Available: https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/

[8] C. Tsallis, S. V. F. Levy, A. M. C. Souza, and R. Maynard, "Statistical-mechanical foundation of the ubiquity of lévy distributions in nature," *Phys. Rev. Lett.*, vol. 75, pp. 3589–3593, Nov 1995. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.75.3589

[9] D. Prato and C. Tsallis, "Nonextensive foundation of lévy distributions," *Phys. Rev. E*, vol. 60, pp. 2398–2401, Aug 1999. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.60.2398

[10] N. Inoue and K. Shinoda, "q-gaussian mixture models for image and video semantic indexing," *Journal of Visual Communication and Image Representation*, vol. 24, no. 8, pp. 1450–1457, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320313001855

[11] D. Ghoshdastidar, A. Dukkipati, and S. Bhatnagar, "q-gaussian based smoothed functional algorithms for stochastic optimization," in *2012 IEEE International Symposium on Information Theory Proceedings*, 2012, pp. 1059–1063. [Online]. Available: https://ieeexplore.ieee.org/document/6283013

[12] R. Tinós and S. Yang, "Use of the q-gaussian mutation in evolutionary algorithms," *Soft Computing*, vol. 15, no. 8, pp. 1523–1549, Aug 2011. [Online]. Available: https://doi.org/10.1007/s00500-010-0686-8

[13] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 1352–1361. [Online]. Available: https://api.semanticscholar.org/CorpusID:11227891

[14] B. D. Ziebart, "Modeling purposeful adaptive behavior with the principle of maximum causal entropy," Ph.D. dissertation, Carnegie Mellon Univ., 2010. [Online]. Available: https://www.cs.cmu.edu/~bziebart/publications/thesis-bziebart.pdf

[15] G. Williams, N. Wagener, B. Goldfain, P. Drews, J. M. Rehg, B. Boots, and E. A. Theodorou, "Information theoretic mpc for model-based reinforcement learning," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1714–1721. [Online]. Available: https://ieeexplore.ieee.org/document/7989202

[16] Z. Wang, O. So, J. Gibson, B. Vlahov, M. S. Gandhi, G.-H. Liu, and E. A. Theodorou, "Variational inference mpc using tsallis divergence," in *Robotics Science and Systems (RSS)*, 2021. [Online]. Available: https://www.roboticsproceedings.org/rss17/p073.pdf

[17] G. Chen, Y. Peng, and M. Zhang, "Effective exploration for deep reinforcement learning via bootstrapped q-ensembles under tsallis entropy regularization," 2018. [Online]. Available: https://arxiv.org/pdf/1809.00403

[18] K. Lee, S. Kim, S. Lim, S. Choi, and S. Oh, "Tsallis reinforcement learning: A unified framework for maximum entropy reinforcement learning," 2019. [Online]. Available: https://arxiv.org/abs/1902.00137

[19] L. Zhu, Z. Chen, E. Uchibe, and T. Matsubara, "Enforcing kl regularization in general tsallis entropy reinforcement learning via advantage learning," 2022. [Online]. Available: https://arxiv.org/abs/2205.07885

[20] D. H. Jacobson and D. Q. Mayne, *Differential dynamic programming*. Elsevier, 1970.

[21] O. So, Z. Wang, and E. A. Theodorou, "Maximum entropy differential dynamic programming," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE Press, 2022, p. 3422–3428. [Online]. Available: https://doi.org/10.1109/ICRA46639.2022.9812228

[22] L.-Z. Liao and C. Shoemaker, "Convergence in unconstrained discrete-time differential dynamic programming," *IEEE Transactions on Automatic Control*, vol. 36, no. 6, pp. 692–706, 1991. [Online]. Available: https://ieeexplore.ieee.org/document/86943

[23] W. J. Thistleton, J. A. Marsh, K. Nelson, and C. Tsallis, "Generalized box–mÜller method for generating q-gaussian random deviates," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4805–4810, 2007. [Online]. Available: https://ieeexplore.ieee.org/document/4385787

[24] C. Vignat and A. Plastino, "Poincaré's observation and the origin of tsallis generalized canonical distributions," *Physica A: Statistical Mechanics and its Applications*, vol. 365, no. 1, pp. 167–172, 2006, fundamental Problems of Modern Statistical Mechanics. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378437106000744

[25] ——, "Central limit theorem and deformed exponentials," *Journal of Physics A: Mathematical and Theoretical*, vol. 40, no. 45, p. F969, oct 2007. [Online]. Available: https://dx.doi.org/10.1088/1751-8113/40/45/F02

[26] Student, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, 1908. [Online]. Available: http://www.jstor.org/stable/2331554

[27] D. Ghoshdastidar, A. Dukkipati, and S. Bhatnagar, "Smoothed functional algorithms for stochastic optimization using q-gaussian distributions," *ACM Trans. Model. Comput. Simul.*, vol. 24, no. 3, jun 2014. [Online]. Available: https://doi.org/10.1145/2628434

[28] T. Luukkonen, "Modelling and control of quadcopter," *Independent research project in applied mathematics, Espoo*, 2011.