2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Adaptive Absolute-Relative Rating for Noise Rejection in Behavioral Cloning based on Tsallis Statistics

Taisuke Kobayashi[1] and Tadayoshi Aoyama[2]

***Abstract*— In robot control from demonstrations, a sufficient dataset cannot be collected for many of the tasks that require experts with qualifications to special skills. Unfortunately, insufficient expert dataset would manifest various types of noise hidden in it. Since adding data is difficult as well, offline imitation learning needs to be robust to such a noise. In the conventional work, a behavioral cloning method based on Tsallis statistics has been developed. However, it weights each data with absolute rating with a fixed threshold, which would fail to imitate coarse/diverse motions. Therefore, this paper improves the conventional method by adding the function of relative rating for each data, which should enable robots to imitate non-noisy data even from coarse/diverse motions. This function can be obtained from a different derivation way of the optimization problem with Tsallis statistics. By integrating it with the conventional derivation way, the proposed method can adjust between the absolute and relative ratings. Finally, for more convenience, we design optimization tricks for the hyperparameters to maximize the variance of weights with avoiding extremely large weights. In numerical simulations and real-robot experiments, we demonstrate the robustness of the proposed method.**

I. Introduction

As the working population declines, there is a growing demand for robots to replace workers for various tasks. This approach is urgently needed because the more specific the skills required for a task, the fewer skilled workers there are (e.g. ICSI by embryologist). For task replacement, imitation learning [1], which reproduces the behavior policy of skilled workers by referring to the data measured during the performance of their tasks, has attracted much attention [2], [3].

However, imitation learning is only practical when there is enough perfect data collected from skilled workers [4], or when there is no risk by operating robots randomly [5]. Both requirements are difficult to meet for tasks that require special skills. The only way is to compromise on the former condition and try to imitate from an insufficient dataset, but in this case, various types of noise become apparent. As a result, the imitation would collapse due to the adverse effects of such a noise.

Therefore, this study focuses on offline learning imitation learning techniques that are robust to noise [6], [7]. These are basically modified versions of the imitation learning technique called behavioral cloning (BC) [4]. By evaluating the validity of each data in the expert dataset, a loss function to be minimized is weighted accordingly. Although these methods have enabled stable imitation by excluding noise even from small datasets, there are still issues in terms of either of applicable condition (i.e. limited to discrete action space) or computational cost. Note that a solver for minimizing the loss function has been made robust recently [8], and since it can be applied complementary to the above approach, we use it toghether with our method in this paper.

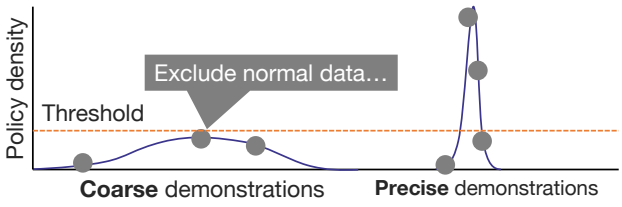

Fig. 1: Issue in the noise-robust BC [9]

To alleviate the issues in the previous methods, a novel noise-robust behavioral cloning has been developed [9], which is derived from Tsallis statistics [10], [11]. This allows imitation while efficiently excluding noise in the dataset with continuous action space. However, the robustness of the method can be interpreted as an implicit weighting based on an *absolute* rating of each data. If the degree of the weighting is not appropriately adjusted by a hyperparameter, the imitation will fail due to the unsuitable weighting. In addition, small weights tend to be assigned to data with coarse/diverse motions, such as initial operation and error recovery, imitating mostly the behaviors in standard situations (see Fig. 1).

To resolve this issue, this paper proposes a method to implicitly weight each data based on its *relative* rating to data in similar situations. For this purpose, we focus on imitation learning as divergence minimization [7], [12], [13]. Specifically, since the usual BC is regarded as a problem of minimizing Kullback-Leibler (KL) divergence between the expert's and imitator's policies, that is extended with Tsallis statistics as the previous method has done so for the log-likelihood maximization problem of BC [9]. In this way, while maximizing the log-likelihood and minimizing the KL divergence are equivalent in the usual BC, they do not coincide when extended to Tsallis statistics. Namely, we derive a different behavior from the conventional method, i.e. the relative rating of each data.

*This work was supported by JST AIP Acceleration Research JPMJCR22U1, Japan.

[1]T. Kobayashi is with the National Institute of Informatics (NII) and with The Graduate University for Advanced Studies (SOKENDAI), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan kobayashi@nii.ac.jp

[2]T. Aoyama is with the Department of Micro-Nano Mechanical Science and Engineering, Graduate School of Engineering, Nagoya University, Nagoya, Aichi, 464-8603, Japan tadayoshi.aoyama@mae.nagoya-u.ac.jp

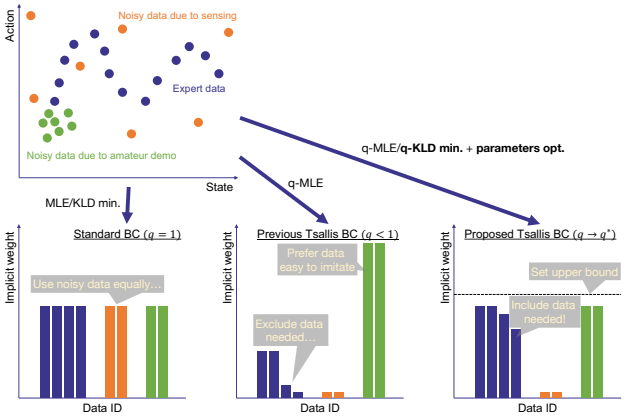

Fig. 2: Improvement of implicit weighting

Since the absolute and relative ratings are expected to have their own pros and cons, we would like to integrate the two methods appropriately. We then integrate them by introducing a new hyperparameter that adjusts the balance of them. In addition, it is desired to tune the hyperparamters (i.e. one to decide the degree of noise rejection given from Tsallis statistics and other to balance the absolute and relative ratings) adaptively according to the given dataset. To this end, a meta-optimization problem is designed: it aims to keep the expected upper bound of the implicit weights at a constant level; and to increase the variance of weights.

The proposed method is validated by numerical simulations. Even when the expert dataset is partially replaced with the artificial noise or amateur dataset, the proposed method succeeds in imitating the expert policy by adjusting the hyperparameters (like Fig. 2). In addition, we demonstrate its robustness in a real-robot page-flipping task.

II. Preliminaries

A. Behavioral cloning

BC [4] is one of the most representative imitation learning methods, which learn a imitator's policy in an offline supervised learning manner. Specifically, under Markov decision process, pairs of state observation $s \in \mathcal{S} \subset \mathbb{R}^{|\mathcal{S}|}$ and action $a \in \mathcal{A} \subset \mathbb{R}^{|\mathcal{A}|}$ are collected by expert(s) with the optimal policy $\pi_{\exp}(a \mid s)$. As a result, an expert dataset $D = \{(s_n, a_n)\}_{n=1}^{N}$ with N pairs are prepared in advance.

With this, the imitator's policy $\pi(a \mid s; \theta)$ with θ the trainable parameters is optimized to match $\pi_{\exp}$ by minimizing the negative log-likelihood as follows:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\theta) \tag{1}$$
$$\mathcal{L}(\theta) = -\mathbb{E}_{(s_n,a_n)\sim D}[\ln \pi(a_n \mid s_n)]$$

As a function approximation of π, we often use neural networks, which has θ mainly as weights and biases parameters, with s as input to output the model parameters of π (e.g. mean μ and scale σ for normal distribution). $\mathcal{L}(\theta)$ is then computed for each $M \leq N$ pairs of batch data, updating θ with backpropagation and stochastic gradietnt decent (e.g. AdaTerm [8] we employed in this paper).

As mentioned in the introduction, the above minimization problem is known to be equivalent to the minimization problem of the KL divergence [7], [12], [13], which is one of the degrees of divergence between probability distributions. In fact, $\mathcal{L}(\theta)$ can be derived by excluding terms independent of θ that are unnecessary for optimization.

$$\begin{aligned}
&\mathbb{E}_{s_n\sim p_e}\left[\mathrm{KL}(\pi_{\exp}(a \mid s_n)||\pi(a \mid s_n;\theta))\right] \\
=&\mathbb{E}_{s_n\sim p_e}\left[\mathbb{E}_{a_n\sim\pi_{\exp}(a|s_n)}\left[-\ln\frac{\ln\pi(a_n \mid s_n;\theta)}{\pi_{\exp}(a_n \mid s_n)}\right]\right] \\
=&-\mathbb{E}_{s_n\sim p_e, a_n\sim\pi_{\exp}(a|s_n)}\left[\ln\pi(a_n \mid s_n;\theta)\right] \\
&-\mathbb{E}_{s_n\sim p_e}\left[H(\pi_{\exp}(a \mid s_n))\right] \\
\propto&-\mathbb{E}_{(s_n,a_n)\sim D}\left[\ln\pi(a_n \mid s_n;\theta)\right] = \mathcal{L}(\theta)
\end{aligned} \tag{2}$$

where, p_e denotes the environmental dynamics, $\mathrm{KL}(p_1||p_2) = \mathbb{E}_{x\sim p_1}[-\ln p_1(x)/p_2(x)]$ defines the KL divergence, and the entropy $H(\cdot)$ is excluded because it does not depend on θ. The expected values for p_e and $\pi_{\exp}$ are approximated by Monte Carlo method with the dataset D actually collected from them. Since the accuracy of this approximation increases with the size of N, it can be seen that BC is better to collect enough data from the expert(s) for more accurate imitation.

B. Tsallis statistics

Tsallis statistics [10], [11] is a system that extends the commonly known Shannon statistics from the underlying information and functions. Here, the natural logarithm is extended to q-logarithm defined below by introducing $q \in \mathbb{R}$.

$$\ln_q(x) = \begin{cases} \ln(x) & q = 1 \\ \dfrac{x^{1-q}-1}{1-q} & q \neq 1 \end{cases} \tag{3}$$

where, $x > 0$ and $\ln_q(x)$ is the monotonic increasing function. With $q \to 1$, it reverts to the natural logarithm; otherwise, it holds $\ln_{q_1}(x) > \ln_{q_2}(x)$ with $q_1 < q_2$ and $x \neq 1$ (if $x = 1$ $\ln_q(x = 1) = 0$ in any q). In addition, it becomes a concave function with $q \geq 0$, and $q < 1$ leads to one-sided boundedness on $x \to 0$: specifically, $\lim_{x\to 0} \ln_{q<1}(x) = -(1-q)^{-1}$.

Other relevant factors of this study are the q-KL divergence (or Tsallis divergence) and the pseudo-additivity. First, the q-KL divergence extends the KL divergence by replacing the natural logarithm in it to q-logarithm, defined as below [14], [15].

$$\mathrm{KL}_q(p_1(x)||p_2(x)) = \mathbb{E}_{x\sim p_1}\left[-\ln_q\frac{p_2(x)}{p_1(x)}\right] \tag{4}$$

where, the probability ratio is given in q-logarithm, which cannot be decomposed into the sums (or differences) of their respective q-logarithms. Instead, the pseudo-additivity yields the following equation [11].

$$\ln_q\frac{p_2}{p_1} = p_1^{q-1}(\ln_q p_2 - \ln_q p_1) \tag{5}$$

Note that this pseudo-additivity prevents the equivalence of the two derivation methods, unlike the standard BC.

C. Generalized behavioral cloning with Tsallis statistics

Using q-logarithm in eq. (3), eq. (1) for minimizing the negative log-likelihood can be extended [9].

$$\theta^* = \arg\min_\theta \mathcal{L}_q(\theta) \tag{6}$$
$$\mathcal{L}_q(\theta) = -\mathbb{E}_{(s_n,a_n)\sim D}[\ln_q \pi(a_n \mid s_n;\theta)]$$

With $q < 1$, $\ln_q(\pi \to 0)$ is bounded as mentioned before. That is, even if noise is mixed in (s_n, a_n), which is unpredictable as $\pi(a_n \mid s_n;\theta) \ll 1$, its gradient is also bounded.

Specifically, the gradient ratio w.r.t. natural logarithm, ρ_q, is given as follows:

$$\rho_q = \frac{\partial \ln_q \pi}{\partial \ln \pi} = \exp\{(1-q)\ln \pi\} \tag{7}$$

As can be seen in this, $\rho_q \to 0$ when $\pi \to 0$ with $q < 1$. In other words, ρ_q corresponds to the implicit weights of data, and if $q < 1$, data with lower likelihood can be weighted closer to zero to suppress their influences. The smaller q is, the larger the change in ρ_q becomes, and the more pronounced the prioritization of data becomes.

In the end, q determines the degree of noise rejection and requires appropriate adjustment. In particular, an extremely small q always causes learning failure because it favors only a small fraction of the data and excludes the others. In addition, as mentioned in the introduction, the implicit weighting by ρ_q depends on the log-likelihood for each data, but this is an absolute rating that does not consider the log-likelihood for other data. It is therefore insufficient for anomaly detection compared to normal data.

III. Formulation

A. Derivation from q-KL divergence

As a preliminary step to derive the proposed method, let's consider the minimization of the q-KL divergence as well as eq. (2) in the usual BC.

$$\begin{aligned}
&\mathbb{E}_{s_n\sim p_e}\left[\mathrm{KL}_q(\pi_{\exp}(a \mid s_n)||\pi(a \mid s_n;\theta))\right] \\
=&\mathbb{E}_{s_n\sim p_e}\left[\mathbb{E}_{a_n\sim\pi_{\exp}(a|s_n)}\left[-\ln_q \frac{\pi(a_n \mid s_n;\theta)}{\pi_{\exp}(a_n \mid s_n)}\right]\right] \\
=&-\mathbb{E}_{s_n\sim p_e, a_n\sim\pi_{\exp}(a|s_n)}\left[\pi_{\exp}^{q-1}(a_n \mid s_n)\ln_q \pi(a_n \mid s_n;\theta)\right] \\
&-\mathbb{E}_{s_n\sim p_e}\left[H_q(\pi_{\exp}(a \mid s_n))\right] \\
\propto&-\mathbb{E}_{(s_n,a_n)\sim D}\left[\pi_{\exp}^{q-1}(a_n \mid s_n)\ln_q \pi(a_n \mid s_n;\theta)\right]
\end{aligned} \tag{8}$$

where, $H_q(\cdot)$ denotes the q-deformed entropy, which is removed due to independence on θ as well as $H(\cdot)$ in eq. (2). We can find that the derived loss function with Monte Carlo approximation is different from eq. (6): namely, whether $\pi_{\exp}^{q-1}(a_n \mid s_n)$ is multiplied or not.

As this term is independent on θ, it can be included in the gradient ratio ρ_q in eq. (7). Using $x^{q-1} = \exp\{-(1-q)\ln(x)\}$, the modified gradient ratio, ρ'_q, is given as follows:

$$\rho'_q = \exp\{(1-q)(\ln \pi - \ln \pi_{\exp})\} \tag{9}$$

Compared to ρ_q, it can be interpreted that each data is weighted by the relative rating of π with $\pi_{\exp}$ as the baseline. This change makes it possible to find relatively important data and prioritize them for imitation. Even if the likelihood is low regardless of noise, for example, when experts operate coarsely without requiring precision, this relative rating can learn the optimal policy while excluding noise.

However, $\pi_{\exp}$ is unknown in practice. That is, if eq. (8) is utilized for imitation, $\pi_{\exp}$ needs to be assumed approximately. As π is learned to imitate $\pi_{\exp}$, we could approximate it by $\pi_{\exp} \simeq \pi$, but ρ'_q would cancel each other out and revert to the usual BC where all data are treated equally.

Instead of π, inspired by the literature [7], we introduce $\bar{\pi}$, which is smoothly updated toward π (and $\pi_{\exp}$ eventually), and assume $\pi_{\exp} \simeq \bar{\pi}$. Specifically, $\bar{\pi}$ is given as a target network, which is a standard technique in deep reinforcement learning [16]. That is, $\bar{\pi}$ has the same network architecture as π with paramters $\bar{\theta}$, which are initialized to be θ and smoothly updated toward θ using CAT-soft update [17].

B. Integration of two loss functions

As shown in eqs. (6) and (8), two different types of loss functions to be minimized were formulated based on Tsallis statistics. The conventional method, eq. (6), does not require any special approximation, while its implicit weight is an absolute rating that depends only on the likelihood (and q) for each data as shown in eq. (7). On the other hand, the new derivation, eq. (8), weights each data relative to $\pi_{\exp}$, enabling the imitator to acquire the optimal behavior in the situations where the likelihood is low overall. However, $\pi_{\exp}$ must be approximated like $\bar{\pi}$ the target network. Since both have their pros and cons, their integration should be useful.

For this integration, we focus on the fact that eq. (6) can be regarded as the special case of eq. (8) with the assumption $\pi_{\exp} = 1$. Accordingly, $\pi_{\exp} = (1-\gamma)+\gamma\bar{\pi}$ is newly assumed with $\gamma \in [0, 1]$ the relative rating level. If γ is small enough, the absolute rating is dominant with the assumption that all data are generated with a certain likelihood; otherwise, the relative rating is activated with $\bar{\pi}$ the baseline. In summary, the proposed method achieves noise-robust imitation learning by solving the following minimization problem.

$$\theta^* = \arg\min_\theta \mathcal{L}_{q,\gamma}(\theta) \tag{10}$$
$$\mathcal{L}_{q,\gamma}(\theta) = -\mathbb{E}_{(s_n,a_n)\sim D}[\pi_\gamma^{q-1}(a_n \mid s_n)\ln_q \pi(a_n \mid s_n;\theta)]$$
$$\pi_\gamma(a_n \mid s_n) = (1-\gamma) + \gamma\bar{\pi}(a_n \mid s_n;\bar{\theta})$$

The gradient ratio in this case, $\rho_{q,\gamma}$, is given as follows:

$$\rho_{q,\gamma} = \exp\{(1-q)(\ln \pi - \ln \pi_\gamma)\} \tag{11}$$

IV. Tuning hyperparameters

Although the loss function to be minimized has been determined as eq. (10), the appropriate q and γ must be given to practically imitate experts while making learning robust to various noises. For this reason, the implicit weight $\rho_{q,\gamma}$ should be varied to improve discrimination performance. At the same time, it is important to avoid making $\rho_{q,\gamma}$ excessively large in order to stabilize learning. To achieve these requirements, the design of γ and optimization of q are proposed below.

A. Design of γ

According to the above two criteria for meta-optimization of q and γ, $\rho_{q,\gamma}$ is restricted to be small relatively. To this end, we first derive the upper bound of $\rho_{q,\gamma}$ in eq. (11) to minimally depend on q and γ as follows:

$$0 \leq \rho_{q,\gamma} = \frac{\pi^{1-q}}{\{(1-\gamma)+\gamma\bar{\pi}\}^{1-q}} \leq \frac{\pi^{1-q}}{(1-\gamma)^{1-q}} \leq \frac{\pi^{1-q}}{1-\gamma} \tag{12}$$

where, $\bar{\pi} \geq 0$ and $(1-\gamma)^{1-q} \geq 1-\gamma$ (since $1-\gamma \leq 1$ is powered by $1-q < 1$) are utilized.

If the expected value of this upper bound is kept at a constant level, the expected value of $\rho_{q,\gamma}$ itself will also be below a constant level, thus enabling stable weighting. Specifically, γ is designed to fix the upper bound at one so that it can also hold for $q \to 1$.

$$\gamma = \max(\min(1 - \mathbb{E}[\pi^{1-q}], \overline{\gamma}), \underline{\gamma}) \tag{13}$$

where, $0 < \underline{\gamma} < \overline{\gamma} < 1$ gives the bounds of γ. The expectation operation is approximated by a sample average per batch of size M selected from the dataset.

As a remark, by setting $\overline{\gamma} < 1$, the upper bound of $\rho_{q,\gamma}$ becomes numerically stable, and by setting $\underline{\gamma} > 0$, we leave room for $\bar{\pi}$ to act. Anyway, as long as γ is not clipped, the expected value of the upper bound of $\rho_{q,\gamma}$ is approximately fixed at one, and $\rho_{q,\gamma}$ is on average less than one. Therefore, it can be expected to stabilize learning and prevent overlearning. The overall smaller weights (and gradients) are not a big issue since the recent stochastic gradient descent method [8] includes gradient scaling.

B. Optimization of q

Next, q is optimized so that $\rho_{q,\gamma}$ varies from data to data. In eq. (11), q is responsible for the scale adjustment of $\rho_{q,\gamma}$. The smaller q is, the more $\rho_{q,\gamma}$ tends to fluctuate, converging to $\rho_{q,\gamma} \to 1$ when $q \to 1$. Therefore, to increase the variance of $\rho_{q,\gamma}$, q needs to be minimized. On the other hand, when γ is clipped in eq. (13), the expected value of the upper bound of $\rho_{q,\gamma}$ deviates from one, and its correction is also required as an equality constraint. In summary, the following minimization problem for q yields the above requirements.

$$\begin{aligned} q^* &= \arg\min_q q \\ \text{s.t. } &\mathbb{E}\left[\frac{\pi^{1-q}}{1-\gamma}\right] = 1,\ \underline{q} < q < 1 \end{aligned} \tag{14}$$

where, $\underline{q} > 0$ gives the explicit lower bound of q because it has been confirmed in the conventional work [9] that learning does not succeed at all for excessively small q.

In practice, the upper and lower bounds of q are forced to be satisfied by using sigmoid function, and the equality constraint on the expected value of the upper bound of $\rho_{q,\gamma}$ is converted to the corresponding regularization term, following Lagrange multiplier method. In summary, we can actually optimize the real number $\phi \in \mathbb{R}$ that replaces q by the following minimization problem.

$$\begin{aligned} \phi^* &= \arg\min_\phi \mathcal{L}(\phi) \\ \mathcal{L}(\phi) &= q(\phi) + \lambda \left| \mathbb{E}\left[\frac{\pi^{1-q(\phi)}}{1-\gamma}\right] - 1 \right| \\ q(\phi) &= \sigma(\phi)(1-\underline{q}) + \underline{q} \end{aligned} \tag{15}$$

where, $\sigma(\cdot) \in (0,1)$ defines the sigmoid function, and $\lambda > 0$ is the Lagrangian multiplier, which adjusts the balance between minimizing q and satisfying the equality constraint. As in eq. (13), the expectation operation is approximated by a batch-wise sample average of size M selected from the dataset and solved by stochastic gradient descent with the main minimization problem in eq. (10). Note that λ can be automatically determined (e.g. [18], [19]), but this paper omits it for simplicity. Although the second term is not needed if γ is not clipped, its gradient is zero in that time and does not affect the update of ϕ (and q), so the branching process is omitted as well.

V. Numerical validation

A. Tasks

As a numerical validation, we employ Ant task, where a quadruped aims to maximize its forward speed, simulated on a Pybullet [20]. A dataset, which has been constructed in the literature [8], is reused. In this dataset, $N = 29,970$ state-action pairs are contained in total.

To evaluate the robustness to noise, two types of noise are prepared. First, artificial noise is generated from a normal distribution $\mathcal{N}(5,1)$, with a probability of 10 % each time, and added to the behavioral data a. The other is to replace one trajectory collected from each expert (i.e. 10 % of the dataset) with a trajectory by an amateur, who converged to a local solution by training. According to the presence/absence of them, four tasks are set.

The imitation performance is evaluated by scores obtained when Ant task is performed using the acquired policies 100 times in total. The score is the median of the sum of the rewards at each time defined in the Ant task itself from the start to the end of the task.

B. Conditions

The imitator's policy π is approximated using neural networks implemented in PyTorch [21]. The network structure consists of two fully connected hidden layers with 100 neurons for each. RMSNorm [22] and Squish function [23], [24] are employed as the combined activation function for the outputs from the hidden layers. Assuming a diagonal normal distribution as the probability model of π, the network outputs the mean and scale of each action dimension.

For optimization, AdaTerm [8], which has been developed for noise-robust optimizer, is employed. The training for each condition is terminated at 200 epochs with the batch size $M = 256$. The random seed used for network initialization, etc., is changed for each trial, and 24 trials are trained under each condition and compared statistically.

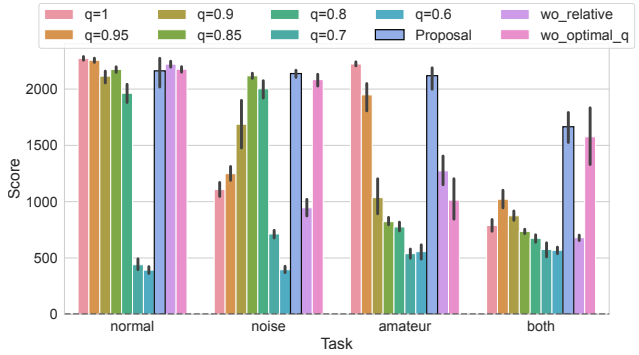

Fig. 3: Imitation scores of Ant task

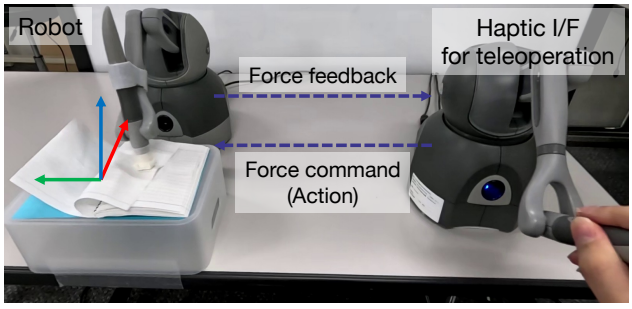

Fig. 4: Setup for page-flipping task

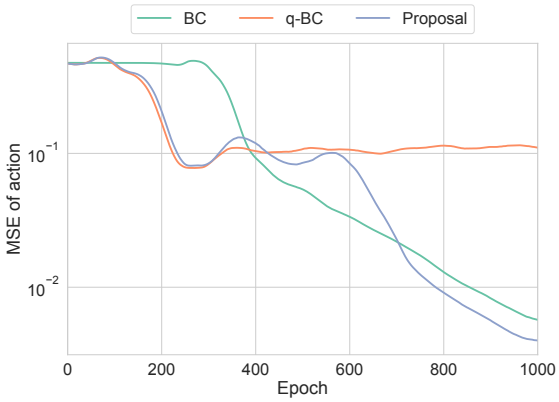

Fig. 5: MSE at page-flipping task

As comparisons, let $q = \{1, 0.95, 0.9, 0.85, 0.8, 0.7, 0.6\}$ be the group of conventional methods including the standard BC (i.e. $q = 1$). As ablation studies, we also provide two types of experiments: one in which the design of γ is excluded from the proposed method and fixed with $\gamma = 0$ (i.e. the conventional method with the optimized q); and another in which q is not optimized and fixed with $q = (1 + \underline{q})/2$ (i.e. the proposed method with the fixed q).

The hyperparameters involved in the proposed method are as follows: $\underline{q} = 0.6$; $\underline{\gamma} = 0.1$; $\overline{\gamma} = 0.9$; and $\lambda = 100$. Note that although there are new hyperparameters that arise to optimize q and γ, all of them have a small dependence on the task to be tackled.

C. Results

The imitation scores with confidence intervals after learning are summarized in Fig. 3. The cases with $q = \{0.6, 0.7\}$ failed to learn in all conditions, which is consistent with the results in the previous study [9]. In other cases, the conventional method was able to accomplish the task under normal conditions without noise, but when the artificial noise was added, the performance was increased as the decrease of q. On the other hand, when replacing with the amateur data, the performance was decreased as q got smaller, which excluded necessary data. These results indicate that the optimal q obviously differs depending on the noise source.

Therefore, when both noises were added, the conventional method no longer works. Even in such a case, the adaptive γ enabled to imitate the expert behaviors robustly, although only it could not solve the case only with amateur noise as the fixed q was not optimal. Hence, the results indicated that the novel loss function of the proposed method and the optimization of the associated hyperparameters are useful for the noise-robust imitation.

VI. Real-robot demonstration

A. Tasks

We demonstrates a page-flipping task, namely, a robot manipulator (3D Systems Touch) tries to flip pages of a ring-bound notebook (see Fig. 4). A demonstration dataset is collected by teleoperating the robot via another one. The system updates the target command at 50 Hz, with the lower feedback loop running at 500 Hz. One trajectory is with 10 seconds, and a total of 60 trajectories (i.e. 10 minutes) were collected. This task has a variety of trajectories because it has degrees of freedom except when flipping pages, and contact forces are not stable, so noise is likely to be mixed in.

The trajectory data includes the end-effector position and exerted force at the current time and one step before as the observed state, and the exerted force after the current time as the action. Since Markov property cannot be satisfied only by this setting, the hidden layers of the network architecture used in the above simulations is replaced by two LSTMs with 100 units. On top of that, to implicitly satisfy Markov property, the next observation is predicted as the literature [25] does, along with the policy. A small amount of noise (i.e. white noise with a scale of 10^{-3}) is also added to the internal state to stabilize the learning, as described in [26].

B. Results

The standard BC (i.e. $q = 1$), the conventional method with $q = 0.8$, labeled q-BC, and the proposed method with the same settings as in the simulations are compared. Their learning curves for the mean squared error (MSE) on the action is depicted in Fig. 5. First, the standard BC did not reduce the error at the beginning of learning due to the difficulty of prediction, while q-BC and the proposed method were able to do it to a certain level by preferentially imitating the relatively easy parts of the data by considering some of the data as noise. However, q-BC did not reduce the error as much as the standard BC and the proposed method because it excluded data that needed to be imitated. Finally, the proposed method obtained the smallest error.

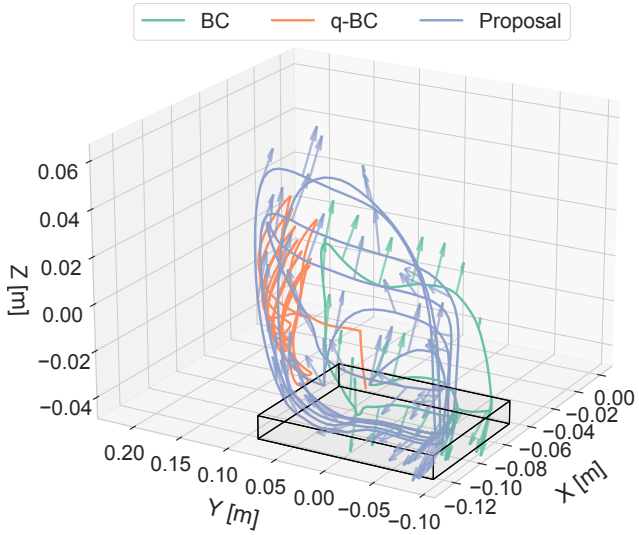

Fig. 6: Typical trajectories with exerted forces as arrows

TABLE I: Performance at page-flipping task

Method	Success rate of continuous flips	Total number of flips
BC	0/10	7
q-BC	0/10	0
Proposal	7/10	34

After learning, the typical trajectories (with exerted forces) for 10 trials with the respective policies are illustrated in Fig. 6. The standard BC flipped the pages by accident, but the end-effector was pressed too hard against the notebook, and it got stuck. In q-BC, the center coordinates of the periodic motion shifted. In contrast, the proposed method flipped the pages with moderate pressure, and then, the end-effector returned to the proper position with a variety of trajectories. The results are summarized in Table I.

VII. Conclusion

This paper developed a new offline imitation learning technique that extends the noise-robust BC based on Tsallis statistics to allow the relative rating, while optimizing the implicit weights for each data to be stable and well distributed. Numerical simulations indicated that the proposed method successfully handles two types of noise. In addition, the robustness of the proposed method was demonstrated in the real-robot page-flipping task.

As a future challenge, since the proposed method is an extension of BC, it inherits its drawbacks such as sensitivity to covariate shifts. Therefore, we aim to integrate the proposed method with some related studies that can alleviate such drawbacks (e.g. [27]) for more practical tasks.

References

[1] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters *et al.*, "An algorithmic perspective on imitation learning," *Foundations and Trends® in Robotics*, vol. 7, no. 1-2, pp. 1–179, 2018.

[2] W. Chi, G. Dagnino, T. M. Kwok, A. Nguyen, D. Kundrat, M. E. Abdelaziz, C. Riga, C. Bicknell, and G.-Z. Yang, "Collaborative robot-assisted endovascular catheterization with generative adversarial imitation learning," in *IEEE International conference on robotics and automation*. IEEE, 2020, pp. 2414–2420.

[3] J. W. Kim, P. Zhang, P. Gehlbach, I. Iordachita, and M. Kobilarov, "Towards autonomous eye surgery by combining deep imitation learning with optimal control," in *Conference on Robot Learning*. PMLR, 2021, pp. 2347–2358.

[4] M. Bain and C. Sammut, "A framework for behavioural cloning." in *Machine Intelligence 15*, 1995, pp. 103–129.

[5] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning," in *International Conference on Machine Learning*, 2000, pp. 663–670.

[6] M. Hussein, B. Crowe, M. Clark-Turner, P. Gesel, M. Petrik, and M. Begum, "Robust behavior cloning with adversarial demonstration detection," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2021, pp. 7835–7841.

[7] F. Sasaki and R. Yamashina, "Behavioral cloning from noisy demonstrations," in *International Conference on Learning Representations*, 2021.

[8] W. E. L. Ilboudo, T. Kobayashi, and T. Matsubara, "Adaterm: Adaptive t-distribution estimated robust moments for noise-robust stochastic gradient optimization," *Neurocomputing*, vol. 557, p. 126692, 2023.

[9] T. Kobayashi and T. Enomoto, "Autonomous driving of personal mobility by imitation learning from small and noisy dataset," in *IEEE/SICE International Symposium on System Integration*. IEEE, 2024, pp. 404–409.

[10] C. Tsallis, "Possible generalization of boltzmann-gibbs statistics," *Journal of statistical physics*, vol. 52, no. 1-2, pp. 479–487, 1988.

[11] H. Suyari and M. Tsukada, "Law of error in tsallis statistics," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 753–757, 2005.

[12] S. K. S. Ghasemipour, R. Zemel, and S. Gu, "A divergence minimization perspective on imitation learning methods," in *Conference on Robot Learning*. PMLR, 2020, pp. 1259–1277.

[13] X. Zhang, Y. Li, Z. Zhang, and Z.-L. Zhang, "f-gail: Learning f-divergence for generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 33, pp. 12 805–12 815, 2020.

[14] F. Nielsen and R. Nock, "A closed-form expression for the sharma–mittal entropy of exponential families," *Journal of Physics A: Mathematical and Theoretical*, vol. 45, no. 3, p. 032003, 2011.

[15] M. Gil, F. Alajaji, and T. Linder, "Rényi divergence measures for commonly used univariate continuous distributions," *Information Sciences*, vol. 249, pp. 124–131, 2013.

[16] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[17] T. Kobayashi, "Consolidated adaptive t-soft update for deep reinforcement learning," in *IEEE World Congress on Computational Intelligence*, 2024.

[18] A. Stooke, J. Achiam, and P. Abbeel, "Responsive safety in reinforcement learning by pid lagrangian methods," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9133–9143.

[19] P. Klink, C. D'Eramo, J. R. Peters, and J. Pajarinen, "Self-paced deep reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9216–9227, 2020.

[20] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," *GitHub repository*, 2016.

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Advances in Neural Information Processing Systems Workshop*, 2017.

[22] B. Zhang and R. Sennrich, "Root mean square layer normalization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[23] J. T. Barron, "Squareplus: A softplus-like algebraic rectifier," *arXiv preprint arXiv:2112.11687*, 2021.

[24] T. Kobayashi and T. Aotani, "Design of restricted normalizing flow towards arbitrary stochastic policy with computational efficiency," *Advanced Robotics*, vol. 37, no. 12, pp. 719–736, 2023.

[25] T. Gangwani, J. Lehman, Q. Liu, and J. Peng, "Learning belief representations for imitation learning in pomdps," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 1061–1071.

[26] S. H. Lim, N. B. Erichson, L. Hodgkinson, and M. W. Mahoney, "Noisy recurrent neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5124–5137, 2021.

[27] K. Brantley, W. Sun, and M. Henaff, "Disagreement-regularized imitation learning," in *International Conference on Learning Representations*, 2019.