

Bayesian Disturbance Injection: Robust Imitation Learning of Flexible Policies

Hanbit Oh[†], Hikaru Sasaki, Brendan Michael and Takamitsu Matsubara¹

Abstract—Scenarios requiring humans to choose from multiple seemingly optimal actions are commonplace, however standard imitation learning often fails to capture this behavior. Instead, an over-reliance on replicating expert actions induces inflexible and unstable policies, leading to poor generalizability in an application. To address the problem, this paper presents the first imitation learning framework that incorporates Bayesian variational inference for learning flexible non-parametric multi-action policies, while simultaneously robustifying the policies against sources of error, by introducing and optimizing disturbances to create a richer demonstration dataset. This combinatorial approach forces the policy to adapt to challenging situations, enabling stable multi-action policies to be learned efficiently. The effectiveness of our proposed method is evaluated through simulations and real-robot experiments for a table-sweep task using the UR3 6-DOF robotic arm. Results show that, through improved flexibility and robustness, the learning performance and control safety are better than comparison methods.

I. INTRODUCTION

Imitation learning provides an attractive means for programming robots to perform complex high-level tasks, by enabling robots to learn skills via observation of an expert demonstrator [1]–[4]. However, two significant limitations of this approach severely restrict applicability to real-world scenarios, these being: 1) learning policies from a human expert demonstrator is often very complex, as humans often choose between multiple optimal actions, and 2) learned policies can be vulnerable to compounding errors during online operation (*i.e.*, there is a lack of robustness). In the former, uncertainties and probabilistic behavior of the human expert (*e.g.*, multiple optimal actions for a task), increases the complexity of learning policies, requiring *flexible policy models*. In the latter case, compounding errors from environmental variations (*e.g.*, task starting position), may induce significant differences between the expert’s training distribution and the applied policy. This drifting issue is commonly referred to as covariate shift [5], and requires *robust policy models*.

To address both these limitations, this paper presents a novel Bayesian imitation learning framework, that learns a probabilistic policy model capable of being both flexible to variations in expert demonstrations, and robust to sources of error in policy application. This is referred to as Bayesian Disturbance Injection (BDI).

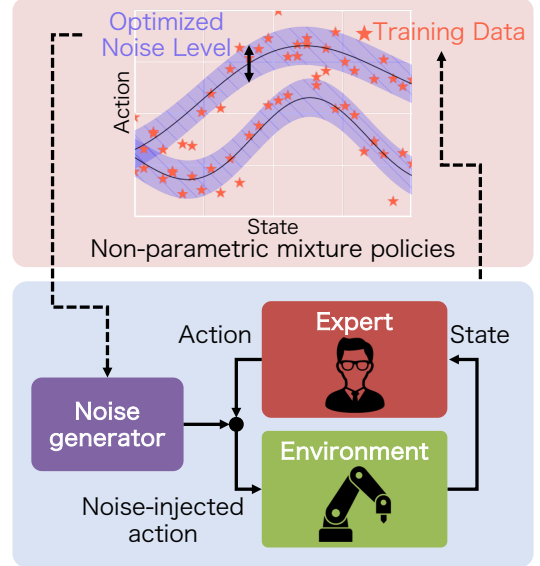


Fig. 1: Overview of multi-modal policy learning with BDI.

Specifically, this paper establishes robust multi-modal policy learning, with flexible regression models [6] as probabilistic non-parametric mixture policies (Fig. 1, top). To induce robustness, noise is injected into the expert’s actions (Fig. 1, bottom) to generate a richer set of demonstrations. Inference of the policy, and optimization of injection noise is performed simultaneously by variational Bayesian learning, thereby minimising covariate shift between the expert demonstrations and learned policy.

To evaluate the effectiveness of the proposed framework, an implementation of Multi-modal Gaussian Process BDI (MGP-BDI) is derived, and experiments in learning probabilistic behavior from a table-sweeping task using the UR3 6-DOF robotic arm is performed. Results show improved flexibility and robustness, with increased learning performance and control safety relative to comparison methods.

II. RELATED WORK

A. Flexibility

A key objective of imitation learning is to ensure that models can capture the variation and stochasticity inherent in human motion. Classical dynamical frameworks for learning trajectories from demonstrations include Dynamic Movement Primitives (DMPs), which can generalize the learned trajectories to new situations (*i.e.*, goal location or speed). However, this generalization depends on the heuristic, thus unsuitable for learning state-dependent feedback policies [7]–[9].

[†] Corresponding author

¹ Authors are with the Division of Information Science, Graduate School of Science and Technology, Nara Institute of Science and Technology (NAIST), Japan. (email: {oh.hanbit.oe9, sasaki.hikaru.rw3, brendan.michael, takam-m}@is.naist.jp)

Gaussian Mixture Regression (GMR) is a non-parametric, intuitive means to learn trajectories or policies from demonstrators in the state-action-space. In this, Gaussian Mixture Modelling (GMM) [10] is used as a basis function to capture non-linearities during learning, and has been utilized in imitation learning that deals with human demonstrations [11]. However, GMR requires to engineer features (e.g., Gaussian initial conditions) by hand to deal with high-dimensional systems. [12].

As an alternative, Gaussian process regression (GPR) deals with implicit (high-dimensional) feature spaces with kernel functions. It thus can directly deal with high-dimensional observations without explicitly learning in this high-dimensional space [13]. In particular, Overlapping Mixtures of Gaussian Processes (OMGP) [14] learns a multi-modal distribution by overlapping multiple GPs, and has been employed as a policy model with multiple optimal actions on flexible task learning of robotic policies [15]. To further reduce a priori tuning, Infinite Overlapping Mixtures of Gaussian Processes (IOMGP) [6] requires only an upper bound of the number of GPs to be estimated. As such, IOMGP is an intuitive means of learning complex multi-modal policies from unlabeled human demonstration data and is employed in this paper.

B. Robustness

While flexibility is key to capturing demonstrator motion, a major issue limiting application of learned policies is the problem of covariate shift. Specifically, environment variations (e.g., manipulator starting position) induces differences between the policy distribution as learned by the manipulator and the actual task distribution during application.

A more general approach to minimizing the covariate shift in imitation learning, is Dataset Aggregation (DAgger) [16], whereby when the robot moves to a state not included in the training data, the expert teaches the optimal recovery. However, this approach has limited applicability in practice due to the risk of running a poorly learned robot and the tediousness to have human experts continue to teach the robot the optimal actions.

An intuitive approach to robustifying learned policies against sources of error, without needing to a priori specify task-relevant learning parameters, is to exploit phenomenon similar to persistence excitation [17]. In this, noise is injected into the expert's demonstrated actions, and the recovery behavior of the expert is learned from this perturbation. In an imitation learning context, Disturbances for Augmenting Robot Trajectories (DART) [18] exploits this phenomenon for learning a deterministic policy model with a single optimal action. Additionally, DART is well suited to creating a richer dataset, by concurrently determining the optimal noise level to be injected into the demonstrated actions during policy learning.

However, the algorithm proposed to achieve DART [18] has a major limitation. It assumes a uni-model deterministic policy, which is unsuitable for many real-world imitation learning tasks, that may consist of multiple optimal actions.

In contrast to that approach, this paper explores a novel method that can eliminate such a severe limitation by incorporating a prior distribution on policy then optimizing both the policy and the injection noise parameters simultaneously via variational Bayesian learning. As such, this novel imitation learning framework improves robustness while maintaining flexibility.

III. PRELIMINARIES

A. Imitation Learning from Expert's Demonstration

The objective of imitation learning is to learn a control policy by imitating the action from the expert's demonstration data. A dynamics model is denoted as Markovian with a state $\mathbf{s}_t \in \mathbb{R}^Q$, an action $a_t \in \mathbb{R}$ and a state transition distribution $p(\mathbf{s}_{t+1} | \mathbf{s}_t, a_t)$. A policy $\pi(a_t | \mathbf{s}_t)$ decides an action from a state. A trajectory $\tau = (\mathbf{s}_0, a_0, \mathbf{s}_1, a_1 \dots a_{T-1}, \mathbf{s}_T)$ which is a sequence of state-action pairs of T steps. The trajectory distribution is indicated as:

$$p(\tau | \pi) = p(\mathbf{s}_0) \prod_{t=0}^{T-1} \pi(a_t | \mathbf{s}_t) p(\mathbf{s}_{t+1} | \mathbf{s}_t, a_t). \quad (1)$$

A key aspect of imitation learning is to replicate the expert's behavior, and as such the function which computes the similarity of two policies using trajectories is defined as:

$$J(\pi, \pi' | \tau) = - \sum_{t=0}^{T-1} \mathbb{E}_{\pi(a|\mathbf{s}_t), \pi'(a'|\mathbf{s}_t)} [||a - a'||_2^2]. \quad (2)$$

A learned policy π^R is obtained by solving the following optimization problem using a trajectory collected by an expert's policy π^* :

$$\pi^R = \arg \max_{\pi} \mathbb{E}_{p(\tau|\pi^*)} [J(\pi, \pi^* | \tau)]. \quad (3)$$

In imitation learning, a learned policy may suffer from the problem of error compounding, caused by covariate shift. This is defined as the distributional difference between the trajectories in data collection, and those in testing:

$$|\mathbb{E}_{p(\tau|\pi^*)} [J(\pi^R, \pi^* | \tau)] - \mathbb{E}_{p(\tau|\pi^R)} [J(\pi^R, \pi^* | \tau)]|. \quad (4)$$

B. Robust Imitation Learning by Injecting Noise into Expert

To learn policies that are robust to covariate shift, DART has previously been proposed [18] for imitation learning problems. In this, expert demonstrations are injected with noise to produce a richer set of demonstrated actions. The level of injection noise is optimized iteratively to reduce covariate shift during data collection.

In this, it is assumed that the injection noise is sampled from a Gaussian distribution as $\epsilon_t \sim \mathcal{N}(0, \sigma_k^2)$, where k is the number of optimizations. The injection noise ϵ_t is added to the expert's action a_t^* .

The noise injected expert's trajectory distribution is denoted as $p(\tau | \pi^*, \sigma_k^2)$ and the trajectory distribution with learned policy as $p(\tau | \pi_k^R)$. To reduce covariate shift,

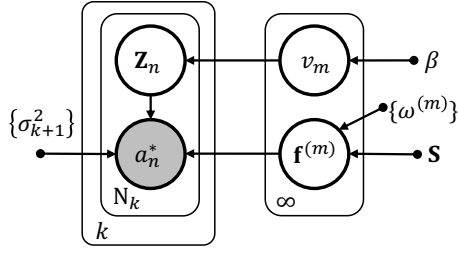


Fig. 2: Graphical model of policy with injection noise parameter

DART introduces upper bound of covariate shift by Pinsker's inequality as:

$$\begin{aligned} & \left| \mathbb{E}_{p(\tau|\pi^*, \sigma_k^2)} [J(\pi^R, \pi^* | \tau)] - \mathbb{E}_{p(\tau|\pi_k^R)} [J(\pi^R, \pi^* | \tau)] \right| \\ & \leq T \sqrt{\frac{1}{2} \text{KL}(p(\tau | \pi_k^R) || p(\tau | \pi^*, \sigma_k^2))}, \end{aligned} \quad (5)$$

where, $\text{KL}(\cdot || \cdot)$ is Kullback-Leibler divergence. However, the upper bound is intractable since the learned policy's trajectory distribution $p(\tau | \pi_k^R)$ is unknown. Therefore, DART solves the upper bound using the noise injected expert's trajectory distribution instead of the learned policy's trajectory distribution. Optimal injection noise distribution is optimized as:

$$\begin{aligned} \sigma_{k+1}^2 &= \arg \max_{\sigma^2} \mathbb{E}_{p(\tau|\pi^*, \sigma_k^2)} \\ & \left[\sum_{t=0}^{T-1} \mathbb{E}_{\pi_k^R(a'_t | \mathbf{s}_t)} [\log \mathcal{N}(a'_t | a_t, \sigma^2)] \right]. \end{aligned} \quad (6)$$

The optimized injection noise distribution can be interpreted as the likelihood of the learned policy.

The learned policy π_k^R used in the optimization of σ_{k+1}^2 is obtained by following:

$$\pi_k^R = \arg \max_{\pi} \sum_{i=1}^{k-1} \mathbb{E}_{p(\tau|\pi^*, \sigma_i^2)} [J(\pi, \pi^* | \tau)]. \quad (7)$$

IV. PROPOSED METHOD

In this section, a novel Bayesian imitation learning framework is proposed (Fig. 1) to learn a multi-modal policy via expert demonstrations with noise injection. Non-parametric mixture model is utilized as a policy prior and incorporates the injection noise distribution as a likelihood to a policy model. An imitation learning method is derived, which learns a multi-modal policy and an injection noise distribution, by variational Bayesian learning. Note, IOMGP [6] is employed as a policy prior in this paper. For simplicity but without loss of generality, this section focuses on one-dimensional action.

A. Policy Model

To learn a multi-modal policy, the policy prior is considered as the product of infinite GPs, inspired by IOMGP. Fig. 2 shows a policy model in which expert's actions $\mathbf{a}^* = [a_n^*]_{n=1}^N$ are estimated by $\mathbf{f}^{(m)}, \mathbf{Z}, \{\sigma_{k+1}^2\}$ where $N = \sum_{j=1}^k N_j$, N_j is a size of the dataset that collected at j -th iteration. The latent function $\mathbf{f}^{(m)}$ is the output of m -th GP given state $\mathbf{S} = [\mathbf{s}_n]_{n=1}^N$. To allocate n -th expert's action a_n^*

to m -th latent function $\mathbf{f}^{(m)}$, indicator matrix $\mathbf{Z} \in \mathbb{R}^{N \times \infty}$ is defined. To estimate the optimal number of GPs, a random variable v_m quantifies the uncertainty assigned to $\mathbf{f}^{(m)}$. In addition, the set of injection noise parameters, $\{\sigma_{k+1}^2\} = \{\sigma_{j+1}^2\}_{j=1}^k$, indicate the distribution of the expert's action a_n^* in the distance from the latent function $\mathbf{f}^{(m)}$.

The set of latent functions is denoted as $\{\mathbf{f}^{(m)}\} = \{\mathbf{f}^{(m)}\}_{m=1}^{\infty}$ and a GP prior is given by :

$$p(\{\mathbf{f}^{(m)}\} | \mathbf{S}, \{\omega^{(m)}\}) = \prod_{m=1}^{\infty} \mathcal{N}(\mathbf{f}^{(m)} | \mathbf{0}, \mathbf{K}^{(m)}; \omega^{(m)}), \quad (8)$$

where $\mathbf{K}^{(m)} = \mathbf{k}^{(m)}(\mathbf{S}, \mathbf{S})$ is the m -th kernel gram matrix with the kernel function $\mathbf{k}^{(m)}(\cdot, \cdot)$ and a kernel hyperparameter $\omega^{(m)}$. Let $\{\omega^{(m)}\} = \{\omega^{(m)}\}_{m=1}^{\infty}$ be the set of hyperparameters of infinite number of kernel functions.

Then the Stick Breaking Process (SBP) [19] is used as a prior of \mathbf{Z} , which can be interpreted as an infinite mixture model as follows:

$$p(\mathbf{Z} | \mathbf{v}) = \prod_{n=1}^N \prod_{m=1}^{\infty} \left(v_m \prod_{j=1}^{m-1} (1 - v_j) \right)^{\mathbf{Z}_{nm}}, \quad (9)$$

$$p(\mathbf{v} | \beta) = \prod_{m=1}^{\infty} \text{Beta}(v_m | 1, \beta). \quad (10)$$

Note that the implementation of variational Bayesian learning cannot deal with infinite-dimensional vectors, so the ∞ component is replaced with a predefined upper bound of the mixtures M . v_m is a random variable indicating the probability that the data corresponds to the m -th GP. Thus, it is possible to estimate the optimal number of GPs with a high probability of allocation starting from an infinite number of GPs. β is a hyperparameter of SBP denoting the level of concentration of the data in the cluster.

The above policy model differs from the IOMGP model for regression [6]; our model employs the set of injection noise parameters $\{\sigma_{k+1}^2\}$ because of the injection noise distribution is updated at each iteration. The next injection noise optimized by (6) deals with the training data collected in the current iteration, while the policy inference (7) deals with all the training data gathered up to the current iteration. Thus, the heteroscedastic Gaussian noise is defined as:

$$\Sigma = \text{diag}\{\sigma_{i+1} \mathbf{1}_{N_i}\}_{i=1}^k, \quad (11)$$

showing the association between the injection noise parameters and the training data collected for each iteration. In this, $\mathbf{1}_{N_i}$ is a vector whose size is N_i and all components are one. In addition, the value of the likelihood (6) does not change even if the mean and the input are replaced, the likelihood for the $\{\mathbf{f}^{(m)}\}, \mathbf{Z}, \{\sigma_{k+1}^2\}$ is derived as :

$$\begin{aligned} & p(\mathbf{a}^* | \{\mathbf{f}^{(m)}\}, \mathbf{Z}; \{\sigma_{k+1}^2\}) \\ &= \prod_{n=1}^N \prod_{m=1}^{\infty} \mathcal{N}(a_n^* | \mathbf{f}_n^{(m)}, \Sigma_{nm})^{\mathbf{Z}_{nm}}. \end{aligned} \quad (12)$$

This formulation is described in a graphical model that defines the relationship between the variables as shown in Fig. 2, and the joint distribution of the model as :

$$\begin{aligned} p(\mathbf{a}^*, \mathbf{Z}, \mathbf{v}, \{\mathbf{f}^{(m)}\} \mid \mathbf{S}; \Omega) \\ = p(\mathbf{a}^* \mid \{\mathbf{f}^{(m)}\}, \mathbf{Z}; \{\sigma_{k+1}^2\}) \cdot \\ p(\{\mathbf{f}^{(m)}\} \mid \mathbf{S}; \{\omega^{(m)}\}) p(\mathbf{Z} \mid \mathbf{v}) p(\mathbf{v} \mid \beta), \end{aligned} \quad (13)$$

where $\Omega = (\{\omega^{(m)}\}, \{\sigma_{k+1}^2\}, \beta)$ represents a set of hyper-parameters.

B. Optimization of Policies and Injection Noise via Variational Bayesian Learning

Bayesian learning is a framework that estimates the posterior distributions of the policies and their predictive distributions for new input data rather than point estimates the policy parameters. To obtain the posterior and predictive distributions, the marginal likelihood is calculated as :

$$\begin{aligned} p(\mathbf{a}^* \mid \mathbf{S}; \Omega) \\ = \iint p(\mathbf{a}^*, \mathbf{Z}, \mathbf{v}, \{\mathbf{f}^{(m)}\} \mid \mathbf{S}; \Omega) d\mathbf{Z} d\mathbf{v} d\{\mathbf{f}^{(m)}\}. \end{aligned} \quad (14)$$

However, it is intractable to calculate the log marginal likelihood of (14) analytically. Therefore, the variational lower bound is derived as the objective function of variational learning. The true posterior distribution is approximated by the variational posterior distribution, which maximizes the variational lower bound. As a common fashion of variational inference, the variational posterior distribution is assumed to be factorized among all latent variables (known as the *mean-field approximation* [20]) as follows:

$$q(\mathbf{Z}, \mathbf{v}, \{\mathbf{f}^{(m)}\}) = q(\mathbf{Z}) \prod_{m=1}^{\infty} q(v_m) q(\mathbf{f}^{(m)}). \quad (15)$$

The variational lower bound $\mathcal{L}(q, \Omega)$ is derived from this assumption by applying the Jensen inequality to the log marginal likelihood, as:

$$\begin{aligned} \log p(\mathbf{a}^* \mid \mathbf{S}; \Omega) \\ \geq \iiint q \log \frac{p(\mathbf{a}^*, \mathbf{Z}, \mathbf{v}, \{\mathbf{f}^{(m)}\} \mid \mathbf{S}; \Omega)}{q} d\mathbf{Z} d\mathbf{v} d\{\mathbf{f}^{(m)}\} \\ = \mathcal{L}(q, \Omega). \end{aligned} \quad (16)$$

In addition, the optimization formulation is derived using the *Expectation-Maximization* (EM) algorithm. The variational posterior distribution q is optimized with a fixed hyperparameter Ω in E-step, and the hyperparameter Ω is optimized with a fixed variational posterior distribution q in M-step with:

$$(\hat{q}, \hat{\Omega}) = \arg \max_{q, \Omega} \mathcal{L}(q, \Omega). \quad (17)$$

See the Section VII-A,B for details of q update laws and lower bound of marginal likelihood. And a summary of the proposed method is shown in Algorithm 1.

Algorithm 1 MGP-BDI

Input: M, K

Output: $q, \hat{\Omega}$

```

1: for  $k = 1$  to  $K$  do
2:   Get dataset through the noise injected expert:
      $\{a_t^*, s_t\}_{t=1}^{N_k} \sim p(\tau \mid \pi^*, \sigma_k^2)$ 
3:   Aggregate datasets  $\mathcal{D} \leftarrow \mathcal{D} \cup \{a_t^*, s_t\}_{t=1}^{N_k}$ 
4:   while  $\mathcal{L}(q, \Omega)$  is not converged do
5:     while  $\mathcal{L}(q, \Omega)$  is not converged do
6:       Update  $q(\mathbf{f}^{(m)})$ ,  $q(\mathbf{Z})$ , and  $q(v_m)$  alternately
7:     end while
8:     Optimize  $\Omega$  with fixed  $q$ :  $\hat{\Omega} \leftarrow \arg \max_{\Omega} \mathcal{L}(q, \Omega)$ 
9:   end while
10: end for
```

C. Predictive Distribution

Using the hyper-parameter Ω and the variational posterior distribution q , optimized by variational Bayesian learning, the predictive distribution of the m -th action $a_*^{(m)}$ and variance $\sigma_*^{2(m)}$ on an current state s_* can be computed analytically as (see [13]).

V. EVALUATION

In this section, the proposed approach is evaluated through simulation and experiments on a 6-DOF robotic manipulator. Specifically, to evaluate the robustness and flexibility of the proposed approach, the following key questions are investigated: i) Does the flexibility induced by MGP-BDI allow for capturing policies with multiple optimal behaviors? ii) Does inducing model robustness via noise injection reduce covariate shift error? iii) Is the optimized injection noise of significantly low variance, allowing for safe perturbations in real-world demonstrations?

To evaluate the proposed method (MGP-BDI), comparisons are made against baseline policy learning methods: 1) Supervised imitation learning (*i.e.*, behavior cloning (BC) [21]) using standard unimodal GPs (UGP-BC) [13], 2) BDI using standard unimodal GPs (UGP-BDI), 3) BC using infinite overlapping mixtures of Gaussian processes (MGP-BC). Specifically, these three are chosen since they represent the state-of-the-art in either flexible or robustness imitation learning. In all experiments, the maximum number of mixture GPs is fixed at ($M = 5$).

A. Simulation

An initial experiment is presented to characterise the flexibility and robustness of the proposed approach, for learning and completing tasks in an environment. Specifically, a standard manipulator learning task (table-sweeping) involving multi-object handling is performed in the V-REP [22] environment, as shown in Fig. 3-(a).

In this experiment, learned policies are evaluated in terms of ability to flexibly learn tasks with multiple optimal actions (*e.g.*, the order in which to sweep the objects from the table), and well as robustness to environmental covariance shift inducing disturbances (*e.g.*, friction between the objects and table inducing variations of object movement).

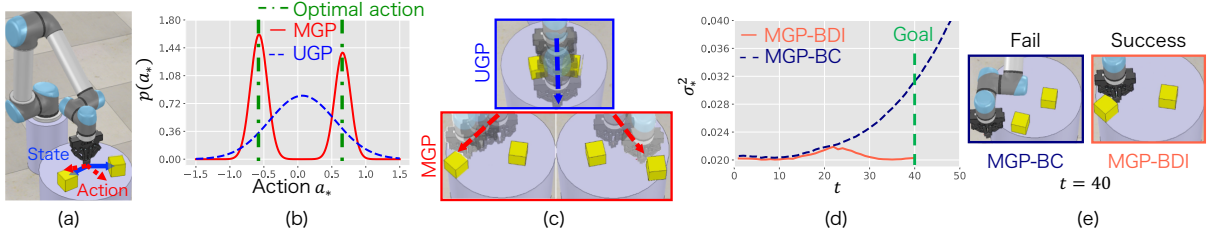


Fig. 3: (a) V-REP environment used in the experiment of the table-sweep task for two boxes. (b), (c): Flexibility comparison between the MGP and UGP. (d), (e): Robustness comparison between the MGP-BDI and MGP-BC.

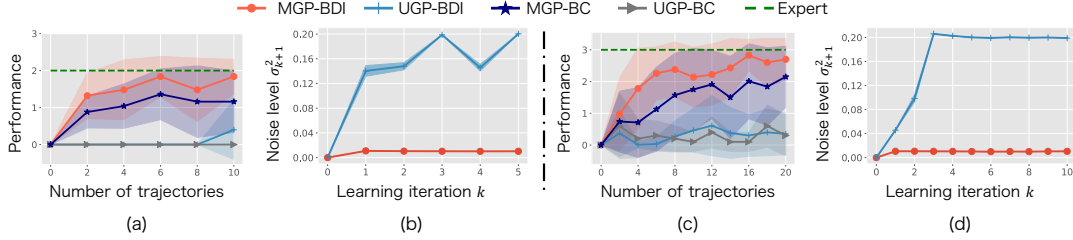


Fig. 4: (a), (c): Comparing learning performance with the number of trajectories. The mean and standard deviation of the performances for 100 test trials. (b), (d): Comparing the tendencies of the injection noise parameters to be optimized according to the training iteration k .

1) *Setup*: Initially, two boxes and the robot arm are placed at fixed coordinates on a table. The state of the system is defined as the relative coordinate between the robot arm and two boxes ($Q = 4$), an action is defined as the velocity of the robot arm in the x and y axis. Demonstrations are generated using a custom PID controller to simulate the human expert, which sweeps the boxes away from the centre. Two demonstrations from these initial conditions are then performed, capturing both variations in the order of which the objects are swept from the table. For each demonstration, the performance is given by the total number of boxes swept off the table (min : 0, max: 2). If a demonstration is unsuccessful (*i.e.*, both boxes were not swept off), the demonstration is discarded, and repeated. Given two successful demonstrations, the data is used to optimize the policy and noise parameters until (17) converges (as seen in Fig. 1). This experiment is repeated for K iterations, appending the successful demonstrations to the training dataset, and continuously updating the policy and noise estimates until convergence of the noise parameters. In the following experiments, $K = 5$ results in convergence. During the test stage, only the robot arm is randomly placed at $s_0 \sim \mathcal{U}(\text{centre} - 0.01\text{m}, \text{centre} + 0.01\text{m})$.

A second experiment is also presented, in a more complex task. To evaluate the proposed method’s scalability, three boxes are placed in the environment, in random coordinates one by one within trisected areas on the table. The state is given by the relative coordinate between the robot arm and three boxes ($Q = 6$), and the performance was calculated as the total number of boxes swept from the table (min : 0, max : 3). Due to the increased task complexity, the maximum number of iterations for updating the policy and noise estimates is $K = 10$.

2) *Result*: The results for these experiments are seen in Fig. 3. In terms of flexibility of learning, Fig. 3-(b) shows that

the unimodal policy learner fails to capture the fact there are multiple optimal actions in the environment, instead learning a mean-centred policy that fails to reach either object (as seen in Fig. 3-(c)). In comparison, the proposed multi-modal approach correctly learns the multi-modal distribution, and outputs actions to sweep the two boxes.

In terms of robustness, it is seen in Fig. 3-(d) that even if a non-parametric flexible policy learner is used (MGP-BC), the variance in the learned action for this policy exponentially increases after the 20-th time-step, and task failure occurs as seen in Fig. 3-(e). This time-step indicates where the robot interacted with the object in the environment, and a sudden increase in uncertainty of the box position occurred. This is a possibility due to the dynamic behavior of the box being different that encountered between this demonstration and the training set, resulting in the model being unable to recover. In comparison, while the proposed noise-injected method also experiments some uncertainty at this interaction, it recovers and retains a near constant certainty throughout the remainder of the task application. In the two-box experiment, MGP-BDI and MGP-BC respectively earned performances $92 \pm 10.7\%$ and $58 \pm 34.7\%$ of the expert demonstrator, and in the three-box experiment this remained similar at $90 \pm 14.3\%$ and $71 \pm 30.9\%$. Other models in which unimodal policies are learned (UGP-BDI and UGP-BC) have been verified to yield about 0 performances.

To evaluate the learning performance, the performance and optimized noise variance is evaluated. In the two-box experiment, Fig. 4-(a), it is seen that the unimodal policy methods (UGP-BC, and UGP-BDI), both fail to learn the multi-modal task, and retain a near zero performance throughout learning. In comparison, the multi-modal policy methods (MGP-BC, and MGP-BDI), both increase learning performance as the number of iterations increase, and MGP-BDI outperforms consistently. In addition to this, it is seen in Fig. 4-(b)

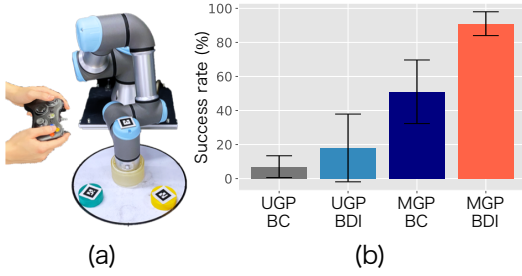


Fig. 5: (a): Experimental environment for 6-DOF robotic arm (UR3) table-sweep task with human expert. (b): Comparison of success rate. Success rate of each learning models is measured over 100 test trials.

when the expert has probabilistic behavior (*i.e.*, multiple targets), the standard UGP-BDI learns an excessive (and potentially unsafe) amount of noise. This is due to the fact that because UGP-BDI assumes a deterministic policy model, a large modeling approximation error is induced, resulting in uncertainty in task performance, and overestimation in the noise variance when attempting to self-correct.

In contrast, the proposed method retains a very low variance in injected noise, by learning policy and noise optimization in a Bayesian framework allowing for confidence in the model to increase gradually as more training data sets are collected. This corresponds to a more stable, and safer demonstration. A similar set of results is seen in the more complex three-box experiment Fig. 4(c-d).

B. Real robot experiment with a human expert

1) *Setup*: The proposed method is evaluated in a real-world robotics task, where the demonstrations of sweeping office supplies (*e.g.*, rubber tape) from a table in the correct order are provided by a human expert, as shown in Fig. 5-(a).

Prior to the start of a demonstration, the two objects are placed at random positions in the upper semicircle of the table. Following the same procedure as outlined in Sec. V-A.1, the human expert performs two demonstrations in which the objects are swept off the table in random order, and the method optimizes the measure and the noise parameter until (17) is converged. This process is repeated $K = 5$ times (10 trajectories). To measure the state of the system (the x,y position of the objects) AR markers are attached to each object (robot arm and rubber tape) and tracked through a RGB-D camera (RealSense D435). As in the previous simulation experiment, the task action is the end-effector velocity. To validate the performance of MGP-BDI in reducing covariate shift, the initial positions of office supplies were arbitrarily placed in a circle on the table to induce scenarios this being a challenging task due to the covariate shift. The learning models' performance is evaluated according to the success of the test episode. Success is determined by the presence of office supplies on the table at the end of the test episode.

2) *Result*: The results of this experiment are seen in Fig. 5-(b). From this, it is seen that the unimodal policy methods (UGP-BC, UGP-BDI) have a poor success rate of $7 \pm 6.4\%$ and $18 \pm 19.9\%$, respectively. As such, they fail to correctly learn policies to account for multiple optimal

actions in the environment, and demonstrate a lack of flexibility. In contrast, the multi-modal policy methods (MGP-BC and MGP-BDI) show improved performance ($51 \pm 18.7\%$ and $91 \pm 7\%$, respectively), however it is clear that even when incorporating flexibility, the success rate for BC is poor.

VI. CONCLUSION

This paper presents a novel Bayesian imitation learning framework, that injects noise into an expert's demonstration, to learn robust multi-optimal policies. This framework captures human probabilistic behavior and allows for learning reduced covariate shift policies, by collecting training data on an optimal set of states. The effectiveness of the proposed method is verified on a real robot with human demonstrations. In the future, this approach will be integrated with kernel approximation methods, or deep neural networks, to learn complex multi-action tasks from unstructured demonstrations (*e.g.*, cooking involving cutting, mixing, and pouring).

VII. APPENDIX

A. Update laws in q

Update of $q(\mathbf{f}^{(m)})$:

$$q(\mathbf{f}^{(m)}) = \mathcal{N}(\mathbf{f}^{(m)} | \mu^{(m)}, \mathbf{C}^{(m)}) \quad (18)$$

$$\mu^{(m)} = \mathbf{C}^{(m)} \mathbf{B}^{(m)} \mathbf{a}^* \quad (19)$$

$$\mathbf{C}^{(m)} = (\mathbf{K}^{(m)-1} + \mathbf{B}^{(m)})^{-1} \quad (20)$$

$$\mathbf{B}^{(m)} = \text{diag}\{r_{nm}/\Sigma_{nn}\} \quad (21)$$

Update of $q(\mathbf{Z})$:

$$q(\mathbf{Z}) = \prod_{n=1}^N \prod_{m=1}^{\infty} r_{nm}^{\mathbf{Z}_{nm}}, r_{nm} = \frac{\rho_{nm}}{\sum_{m=1}^{\infty} \rho_{nm}} \quad (22)$$

$$\log \rho_{nm} = -\frac{1}{2} \log(2\pi \Sigma_{nn}) - \frac{1}{2\Sigma_{nn}} (\mathbf{a}_n^* - \mu_n^{(m)})^2 + \psi(\alpha_m) - \psi(\alpha_m + \beta_m) + \sum_{j=1}^{m-1} \{\psi(\beta_j) - \psi(\alpha_j + \beta_j)\} \quad (23)$$

where, $\psi(\cdot)$ is the digamma function.

Update of $q(v_m)$:

$$q(v_m) = \text{Beta}(v_m | \alpha_m, \beta_m) \quad (24)$$

$$\alpha_m = 1 + \sum_{n=1}^N r_{nm}, \beta_m = \beta + \sum_{j=m+1}^{\infty} \sum_{n=1}^N r_{nj} \quad (25)$$

B. Lower bound of marginal likelihood:

$$\begin{aligned} \mathcal{L}(q, \Omega) &= \sum_{m=1}^{\infty} \log \mathcal{N}(\mathbf{a}^* | \mathbf{0}, \mathbf{K}^{(m)} + \mathbf{B}^{(m)-1}) - \text{KL}(q(\mathbf{v}) || p(\mathbf{v})) \\ &+ \sum_{n=1}^N \sum_{m=1}^{\infty} r_{nm} \left\{ -\frac{1}{2} \log(2\pi \Sigma_{nn}) + \psi(\alpha_m) - \psi(\alpha_m + \beta_m) \right. \\ &\quad \left. + \sum_{j=1}^{m-1} \{\psi(\beta_j) - \psi(\alpha_j + \beta_j)\} - \log r_{nm} \right\} \end{aligned} \quad (26)$$

REFERENCES

- [1] A. Coates, P. Abbeel, and A. Y. Ng, “Apprenticeship learning for helicopter control,” *Communications of the ACM*, vol. 52, no. 7, pp. 97–105, 2009.
- [2] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, *et al.*, “End to end learning for self-driving cars,” *Neural Information Processing Systems (NIPS). Deep Learning Symposium*, 2016.
- [3] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel, “Deep imitation learning for complex manipulation tasks from virtual reality teleoperation,” in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 1–8.
- [4] 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.
- [5] S. Ross and D. Bagnell, “Efficient reductions for imitation learning,” in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 661–668.
- [6] J. Ross and J. Dy, “Nonparametric mixture of gaussian processes with constraints,” in *International Conference on Machine Learning (ICML)*, 2013, pp. 1346–1354.
- [7] S. Schaal, J. Peters, J. Nakanishi, and A. Ijspeert, “Learning movement primitives,” in *Robotics research. the eleventh international symposium*. Springer, 2005, pp. 561–572.
- [8] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, “Dynamical movement primitives: learning attractor models for motor behaviors,” *Neural computation*, vol. 25, no. 2, pp. 328–373, 2013.
- [9] S. M. Khansari-Zadeh and A. Billard, “Learning stable nonlinear dynamical systems with gaussian mixture models,” *IEEE Transactions on Robotics*, vol. 27, no. 5, pp. 943–957, 2011.
- [10] S. Calinon, “A tutorial on task-parameterized movement learning and retrieval,” *Intelligent service robotics*, vol. 9, no. 1, pp. 1–29, 2016.
- [11] M. Kyrarini, M. A. Haseeb, D. Ristić-Durrant, and A. Gräser, “Robot learning of industrial assembly task via human demonstrations,” *Autonomous Robots*, vol. 43, no. 1, pp. 239–257, 2019.
- [12] Y. Huang, L. Roza, J. Silvério, and D. G. Caldwell, “Kernelized movement primitives,” *The International Journal of Robotics Research*, vol. 38, no. 7, pp. 833–852, 2019.
- [13] C. E. Rasmussen, “Gaussian processes in machine learning,” in *Summer School on Machine Learning*. Springer, 2003, pp. 63–71.
- [14] M. Lázaro-Gredilla, S. Van Vaerenbergh, and N. D. Lawrence, “Overlapping mixtures of gaussian processes for the data association problem,” *Pattern recognition*, vol. 45, no. 4, pp. 1386–1395, 2012.
- [15] H. Sasaki and T. Matsubara, “Multimodal policy search using overlapping mixtures of sparse gaussian process prior,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 2433–2439.
- [16] S. Ross, G. Gordon, and D. Bagnell, “A reduction of imitation learning and structured prediction to no-regret online learning,” in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 627–635.
- [17] S. Sastry and M. Bodson, *Adaptive control: stability, convergence and robustness*. Courier Corporation, 2011.
- [18] M. Laskey, J. Lee, R. Fox, A. Dragan, and K. Goldberg, “DART: Noise injection for robust imitation learning,” in *Conference on Robot Learning (CoRL)*, 2017, pp. 143–156.
- [19] J. Sethuraman, “A constructive definition of dirichlet priors,” *Statistica sinica*, pp. 639–650, 1994.
- [20] G. Parisi, *Statistical field theory*. Addison-Wesley, 1988.
- [21] M. Bain and C. Sammut, “A framework for behavioural cloning,” in *Machine Intelligence 15*, 1995, pp. 103–129.
- [22] E. Rohmer, S. P. Singh, and M. Freese, “V-REP: A versatile and scalable robot simulation framework,” in *International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1321–1326.