Anand, Akhil S.; Kaushik, Rituraj; Gravdahl, Jan Tommy; Abu-Dakka, Fares J.

Data-efficient Reinforcement Learning for Variable Impedance Control

Published in:
IEEE Access

DOI:
10.1109/ACCESS.2024.3355311

Published: 01/01/2024

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY-NC-ND

Please cite the original version:
Anand, A. S., Kaushik, R., Gravdahl, J. T., & Abu-Dakka, F. J. (2024). Data-efficient Reinforcement Learning for Variable Impedance Control. *IEEE Access*, *12*, 15631-15641. https://doi.org/10.1109/ACCESS.2024.3355311

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 December 2023, accepted 8 January 2024, date of publication 17 January 2024, date of current version 1 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3355311

RESEARCH ARTICLE

Data-Efficient Reinforcement Learning for Variable Impedance Control

AKHIL S. ANAND[1], (Member, IEEE), RITURAJ KAUSHIK[2], JAN TOMMY GRAVDAHL[1], (Senior Member, IEEE), AND FARES J. ABU-DAKKA[3], (Member, IEEE)

[1]Department of Engineering Cybernetics, Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway
[2]Intelligent Robotics Group, Department of Electrical Engineering and Automation (EEA), Aalto University, 00076 Espoo, Finland
[3]Department of Electronic and Informatics, Faculty of Engineering, Mondragon University, Arrasate, 20500 Mondragon, Spain

Corresponding author: Akhil S. Anand (akhil.s.anand@ntnu.no)

This work was supported in part by the project ''Dynamic Robot Interaction and Motion Compensation'' funded by the Research Council of Norway under Contract 270941, and in part by the Basque Government (ELKARTEK) projects under Grant Proflow KK-2022/00024 and Grant HELDU KK-2023/00055.

ABSTRACT One of the most crucial steps toward achieving human-like manipulation skills in robots is to incorporate compliance into the robot controller. Compliance not only makes the robot's behaviour safe but also makes it more energy efficient. In this direction, the variable impedance control (VIC) approach provides a framework for a robot to adapt its compliance during execution by employing an adaptive impedance law. Nevertheless, autonomously adapting the compliance profile as demanded by the task remains a challenging problem to be solved in practice. In this work, we introduce a reinforcement learning (RL)-based approach called DEVILC (Data-Efficient Variable Impedance Learning Controller) to learn the variable impedance controller through real-world interaction of the robot. More concretely, we use a model-based RL approach in which, after every interaction, the robot iteratively learns a probabilistic model of its dynamics using the Gaussian process regression model. The model is then used to optimize a neural-network policy that modulates the robot's impedance such that the long-term reward for the task is maximized. Thanks to the model-based RL framework, DEVILC allows a robot to learn the VIC policy with only a few interactions, making it practical for real-world applications. In simulations and experiments, we evaluate DEVILC on a Franka Emika Panda robotic manipulator for different manipulation tasks in the Cartesian space. The results show that DEVILC is a promising direction toward autonomously learning compliant manipulation skills directly in the real world through interactions. A video of the experiments is available in the link: https://youtu.be/_uyr0Vye5no.

INDEX TERMS Model-based reinforcement learning, variable impedance learning control, Gaussian processes, covariance matrix adaptation.

I. INTRODUCTION

Robot based automation technology promises to solve many critical real-world applications in industries, healthcare to households in the near future. Robots with the capability to manipulate objects with human-level dexterity is a key aspect for such applications, especially in unstructured environments. While interacting with an unstructured environment, robots have to compensate for uncertainties in the interaction dynamics through the control law. Adapting muscle stiffness and, therefore, arm impedance has been proposed as a mechanism by which humans achieve compliance, which is central to our dexterity [1], [2], [3]. Drawing motivation from human manipulation skills, incorporating compliance behavior into robot control has been investigated as a way to achieve safe and dexterous manipulation skills. Impedance Control (IC) was introduced as a result of this approach and was successfully demonstrated in robotic manipulation [4] by

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Alshabi.

incorporating compliance behavior in robot position or force control.

Traditionally, robotic manipulators have relied on position control, which does not ensure safety and energy efficiency during constrained interactions. The IC, on the other hand, offers a framework to achieve compliant manipulation skills with safety guarantees and energy efficiency. Unlike conventional control approaches, IC models a dynamic relation between manipulator variables such as the end-point position and force rather than controlling these variables independently. IC could provide a feasible solution to overcome position uncertainties and avoid large impact forces since robots are controlled to modulate their motion or compliance based on the sensed forces as feedback [5]. IC is naturally extended to Variable Impedance Control (VIC) where the impedance parameters are varied during the task [6]. VIC gained popularity in robotic research due to its adaptability and safety properties.

With the advancement of learning-based techniques in the field of robotics and control, they have become increasingly popular in the domain of variable impedance control, also known as Variable Impedance Learning Control (VILC) [7]. Several approaches, including Imitation Learning (IL), inverse reinforcement learning (IRL), and Reinforcement Learning (RL), are utilized in VILC. Out of all these methods, RL has the capability to learn intricate and sophisticated control policies, but at the cost of requiring a vast amount of data obtained through extended interactions between the robot and its environment. To enhance the data efficiency of RL, Model-based Reinforcement Learning (MBRL) methods offer a promising solution by using a progressively learned dynamical model of both the robot and its environment [8], [9].

Recently, Probabilistic Inference for Learning Control (PILCO) [8], a highly data-efficient MBRL approach has been applied to VILC [10]. However, PILCO imposes limitations on the reward functions and policy structure, limiting the use of arbitrary rewards as in a standard RL setting. Moreover, analytical approaches like PILCO are not amenable to efficient parallelization on multi-core computers [9]. To address these limitations, an alternative approach called Black-DROPS [9] has been proposed which is based on the gradient-free black-box search algorithm Covariance Matrix Adaptation (CMA-ES) [11]. This approach imposes no constraints on reward functions or policies and can be easily parallelized. By combining Gaussian Processes (GP) based dynamics learning with CMA-ES based policy optimization, a highly data-efficient MBRL framework is achieved, taking into account the uncertainty in the model and performing a comprehensive policy search.

In this work, we focus on learning an optimal impedance adaptation strategy for a VIC in the context of robotic manipulation through real-world interaction of the robot. We take inspiration from Black-DROPS framework and propose a VILC framework called DEVILC (Data-Efficient Variable Impedance Learning Controller) that is highly data-efficient and does not impose any restrictions on the structures of the policy or the reward function. In addition, unlike PILCO-based VILC, DEVILC can make use of multi-core processors to parallelize the policy optimization process, making it suitable for real-world application.

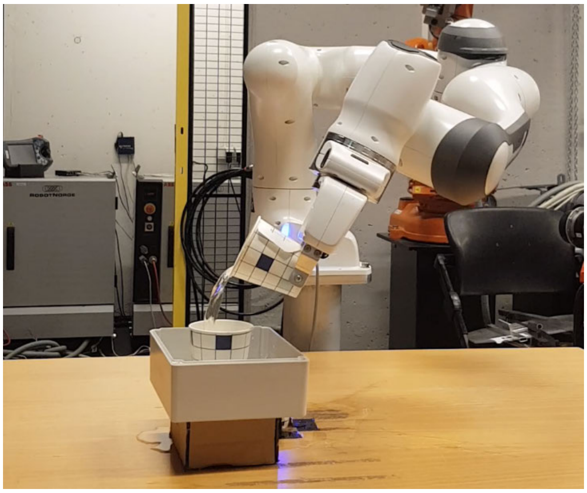


FIGURE 1. Experimental setup with Franka-Emika Panda robotic manipulator pouring water into a cup.

In summary, the main contributions of this paper are:

- We propose a model-based VILC framework called DEVILC using GP models and using the evolution strategy, CMA-ES to optimize a Neural Network (NN) policy.
- We demonstrate that DEVILC is highly data-efficient for learning impedance adaptation strategy for robotic manipulation.

The structure of the rest of this paper is as follows. In Section II, we review existing literature relevant to our work. Section III provides a brief overview of the necessary background knowledge. In Section IV, we present the details of our proposed model-based reinforcement learning (MBRL) based variable impedance learning control (VILC) framework. Section V evaluates the proposed VILC framework through simulations and experiments using a Franka Panda robotic manipulator. Finally, we provide a detailed discussion on the results and conclusions in Section VI and Section VII respectively.

II. RELATED WORK

A wide variety of learning-based approaches are combined with VIC to develop various VILC methods [7]. Prominent examples of such learning-based approaches are IL, Iterative learning control (ILC), and RL. IL has been used in many recent VILC works [12], [13], [14], [15]. IL-based VILC methods are generally some form of Learning from Demonstration (LfD) methods as they often rely on demonstrations to learn from [16]. IL can be useful in developing highly sample efficient VILC [7]. But such learning strategies can be biased to the demonstration which are often suboptimal and potentially limit the performance and generalization of the learned policies. IL is useful for tasks that are easy to demonstrate and which do not have a clear optimal way of execution, whereas RL is well suited for highly-dynamic tasks, where there is a clear measure of the success of the task [17]. Optimizing variable impedance gains/parameters can be done using ILC where the robot improves its performance iteratively. ILC based methods have been used for VILC in a range of works [18], [19], [20], [21]. The key difference between ILC and RL is that, in RL, the control law is derived by maximizing a reward function defined by the task requirements. One advantage of ILC compared to RL is its sample efficiency. But even when a model of the dynamics is not available, RL offers better performance and can be applied to a broader range of problems [22]. In the rest of this section, we provide a brief review of the related works relevant to this paper covering the area of VILC. In our discussions, we focus on RL-based and data-efficient VILC approaches.

A. RL-BASED VILC

Recently, RL has been widely explored for VILC research. Several previous works used deep-RL to learn VILC for various robotic manipulation tasks [23], [24], [25], [26], [27]. However, RL demands a large amount of data samples through interactions for learning policies. Authors of [23] compare different action spaces in deep-RL for robotic manipulation. In [24], an RL framework for learning contact-rich manipulation tasks is proposed where an off-policy model-free RL algorithm (SAC) is used to learn the stiffness and position parameters of a parallel position-force controller. This approach could learn policies achieving high success rates in insertion tasks even under uncertainties. In [25], the approach was extended to real robotic manipulators for safely learning contact-rich manipulation tasks. A parallel position/force control and admittance control were evaluated in this framework on position-controlled robots. Similar to other approaches, in [26], the right choice of action space for learning contact-rich tasks in presence of uncertainties was investigated. A model-free RL approach is used to learn policies for direct torque control, fixed gain PD control, and variable gain PD control. On comparing the three controllers, the variable gain PD controlled demonstrated superior performance and reliability. Similarly in [27], on comparing direct torque control, joint PD, inverse dynamics, and task-space impedance control, the impedance control showed superior performance compared to the other three. But all of these approaches have the drawback of model-free RL in terms of low data efficiency and lack of task transferability.

B. DATA-EFFICIENT VILC

All approaches mentioned above could learn complex VIC policies for specific tasks, however, at the expense of data efficiency. Authors in [28] demonstrated model-free RL based VILC using Dynamic Movement Primitive (DMP) policy [29] and Policy Improvement with Path Integrals algorithm (PI^2) [30], which is data-efficient but it fails to scale to complex policies. In [31], PI^2 approach was used to learn torque control profiles for robot manipulators for compliant manipulation using desired position trajectories from kinesthetic demonstrations. But it is not suitable for force-based VIC, as unlike stiffness values the impedance parameters can not be estimated directly from kinesthetic demonstrations used in [32]. Kim et al. [32] demonstrated that augmenting position demonstrations with stiffness estimates and using it to learn a stiffness controller could provide superior manipulation performance compared to a position controller. Alternatively, MBRL approaches offer a data-efficient and scalable framework leveraging on a learned dynamical model. In [33], MBRL is used to learn position-based VIC on industrial robots using GP models. Authors of [10] and [34] used a similar approach for force-based VIC and hybrid force-motion control for contact-sensitive tasks. In [35], Probabilistic Ensembles with Trajectory Sampling (PETS) approach is used to learn a position-based VIC strategy for Human-Robot Collaboration (HRC) tasks. Although MBRL methods are data-efficient compared to model-free RL, it still demands a lot of interactions, making it difficult to apply to robotic manipulation tasks. In this work, we tackle this issue by using a micro-data-based policy optimization method combining GP models and CMA-ES based policy optimization [9].

TABLE 1. List of variables.

Symbol	Description
$\mathbf{q}, \dot{\mathbf{q}}$	Joint positions and velocity vectors
$\mathbf{x}, \dot{\mathbf{x}}, \ddot{\mathbf{x}}$	End-effector position, velocity and acceleration in task space
$\mathbf{x^r}, \dot{\mathbf{x}}^\mathbf{r}, \ddot{\mathbf{x}}^\mathbf{r}$	Reference/desired end-effector position, velocity and acceleration in task space
$\mathbf{f_c}$	Task space control force
$\mathbf{f_{ext}}$	External force acting on the end-effector
$\mathbf{\Lambda}(\mathbf{q})$	Cartesian inertia matrix
$\mathbf{\Gamma}(\mathbf{q}, \dot{\mathbf{q}})$	Matrix for centrifugal and Coriolis effects
$\boldsymbol{\eta}(\mathbf{q})$	Gravitational force
$\mathbf{J}$	End-effector geometric Jacobian
$\mathbf{H}(\mathbf{q})$	Joint space inertia matrix
$\mathbf{V}(\mathbf{q}, \dot{\mathbf{q}})$	Centrifugal and Coriolis matrices in joint space
$\mathbf{M}, \mathbf{D}, \mathbf{K}$	Impedance matrices (mass, damping, stiffness)
$\boldsymbol{\pi}(\mathbf{s}_t, \mathbf{f}^t\mathbf{_{ext}} \mid \boldsymbol{\theta})$	Impedance adaptation policy
$\boldsymbol{\theta}$	Policy parameters

An alternative approach in [36], proposed an Model Predictive Control (MPC) based VIC framework termed as deep model predictive variable impedance control. This approach provides a data-efficient VILC framework which is scalable to complex tasks and easily transferable across different task scenarios. But this framework still demands a considerable amount of training data to learn a quality model as it uses an ensemble of probabilistic neural networks to model the generalized Cartesian impedance dynamics of the robot. Additionally, it demands high computational time due to the sampling-based MPC scheme used for

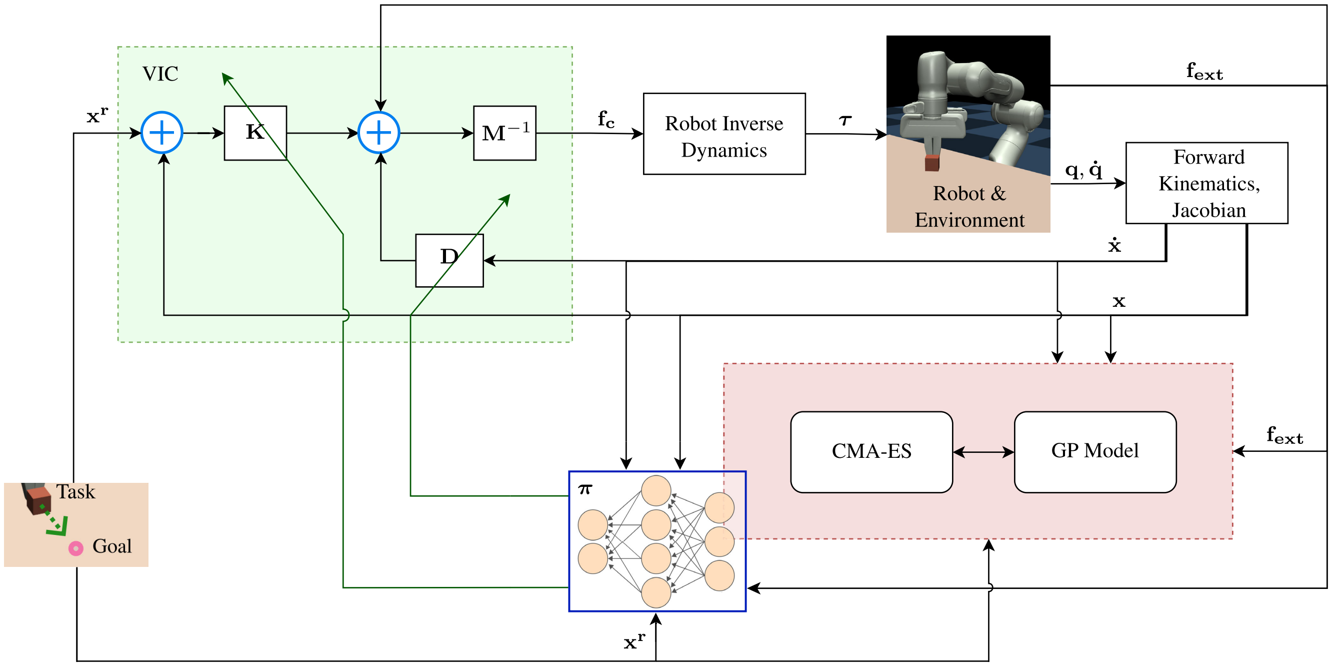


FIGURE 2. DEVILC framework with GP dynamics model and CMA-ES based policy optimization to learn an NN based impedance adaptation policy. The task objective is given by (7).

impedance optimization, therefore limiting the impedance adaptation frequency in real-time. This approach is more relevant in cases where it is required to learn a policy that can be transferred to different tasks or task scenarios and does not demand high-frequency impedance adaptation. The proposed Data-Efficient Variable Impedance Learning Controller (DEVILC) approach learns a task-specific policy, therefore the learned policy need not be easily transferable to another task. But it is more data-efficient owing to GP based dynamics model. Additionally, the DEVILC approach does not have any limitation on impedance adaptation frequency primarily due the decision making component of the approach (NN policy in this paper) which is cheap to compute in real-time.

III. BACKGROUND

A. ROBOT MANIPULATOR DYNAMICS

For a rigid n-DOF robotic arm, the task space formulation of robot dynamics is given by

$$\mathbf{\Lambda}(\mathbf{q})\ddot{\mathbf{x}} + \mathbf{\Gamma}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{x}} + \eta(\mathbf{q}) = \mathbf{f_c} - \mathbf{f_{ext}}\,, \tag{1}$$

where $\dot{\mathbf{x}}, \ddot{\mathbf{x}}$ are the velocity and acceleration of the robot end-effector in task space. $\mathbf{f_c}$ is the task space control force, $\mathbf{f_{ext}}$ is the external force, $\mathbf{\Gamma}(\mathbf{q}, \dot{\mathbf{q}}) \in \mathbb{R}^{\mathbf{6}\times\mathbf{6}}$ is matrix representing the centrifugal and Coriolis effects, and $\eta(\mathbf{q}) = \mathbf{J^{-T}}\mathbf{g}(\mathbf{q}) \in \mathbb{R}^{\mathbf{6}\times\mathbf{1}}$ is the gravitational force, where $\mathbf{g}(\mathbf{q})$ is the joint space forces and torques. The Cartesian inertia matrix is denoted as $\mathbf{\Lambda}(\mathbf{q}) = (\mathbf{J}\mathbf{H}(\mathbf{q})^{-\mathbf{1}}\mathbf{J^T})^{-\mathbf{1}} \in \mathbb{R}^{\mathbf{6}\times\mathbf{6}}$, where $\mathbf{H}(\mathbf{q}) \in \mathbb{R}^{n\times n}$ is the joint space inertia matrix and $\mathbf{J}$ is the end-effector geometric Jacobian. By additionally knowing the centrifugal and Coriolis matrices in joint space, $\mathbf{V}(\mathbf{q}, \dot{\mathbf{q}})$, the corresponding task space matrix is given by

$$\mathbf{\Gamma}(\mathbf{q}, \dot{\mathbf{q}}) = \mathbf{J}^{-T}\mathbf{V}(\mathbf{q}, \dot{\mathbf{q}})\mathbf{J}^{-1} - \mathbf{\Lambda}(\mathbf{q})\dot{\mathbf{J}}\mathbf{J}^{-1}\,. \tag{2}$$

B. VARIABLE IMPEDANCE CONTROL

VIC is designed to achieve force regulation by adjusting the system impedance [37], via the adaptation of the inertia, damping, and stiffness matrices. In the presence of a force and torque sensor measuring f_{ext}, impedance control can be implemented by enabling inertia shaping [38]. Casting the control law

$$\mathbf{f_c} = \mathbf{\Lambda}(\mathbf{q})\boldsymbol{\alpha} + \mathbf{\Gamma}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{x}} + \eta(\mathbf{q}) + \mathbf{f_{ext}}\,, \tag{3}$$

into the dynamic model in (1) results in $\ddot{\mathbf{x}} = \boldsymbol{\alpha}$, $\boldsymbol{\alpha}$ being the control input that denotes acceleration with respect to the base frame. In task space IC, the objective is to maintain a dynamic relationship (4) between the external force, $\mathbf{f_{ext}}$, and the error in position $\boldsymbol{\delta}\mathbf{x} = \mathbf{x^r} - \mathbf{x}$, velocity $\delta\dot{\mathbf{x}} = \dot{\mathbf{x}}^r - \dot{\mathbf{x}}$ and acceleration $\delta\ddot{\mathbf{x}} = \ddot{\mathbf{x}}^r - \ddot{\mathbf{x}}$. This dynamic relationship that governs the interaction is modeled as a mass-spring-damper system as follows

$$\mathbf{M}\delta\ddot{\mathbf{x}} + \mathbf{D}\delta\dot{\mathbf{x}} + \mathbf{K}\delta\mathbf{x} = \mathbf{f_{ext}}\,, \tag{4}$$

where $\mathbf{M}$, $\mathbf{D}$ and $\mathbf{K}$ are Symmetric Positive Definite (SPD) matrices, adjustable impedance parameters, representing inertia, damping and stiffness terms, respectively. This desired dynamic behavior (4) can be achieved using the following control law,

$$\boldsymbol{\alpha} = \ddot{\mathbf{x}}^r + \mathbf{M}^{-\mathbf{1}}(\mathbf{D}\delta\dot{\mathbf{x}} + \mathbf{K}\delta\mathbf{x} - \mathbf{f_{ext}})\,. \tag{5}$$

Without external force acting on the manipulator, the end-effector will asymptotically follow the desired trajectory under this control scheme. In the presence of external forces, the compliant behavior of the end-effector is described by (4).

C. CMA-ES

The Covariance Matrix Adaptation (CMA-ES) is an evolutionary strategy designed to solve non-convex and non-linear black-box optimization problems in continuous domain [11]. It is one of the state-of-the-art methods in evolutionary computation, specifically for continuous optimization. CMA-ES uses a multivariate Gaussian distribution $\mathcal{N}\left(\boldsymbol{\mu}^{\mathbf{d}}, \Sigma^{\mathbf{d}}\right)$ where $\boldsymbol{\mu}^{\mathbf{d}} \in \mathbb{R}^{d}$, $\Sigma^{\mathbf{d}} \in \mathbb{R}^{d \times d}$ is a positive definite symmetric matrix and d is the dimension of a solution vector. CMA-ES algorithm works by (i) drawing β candidate samples following the multivariate Gaussian distribution, (ii) evaluating these β samples based on a given cost function, (iii) selecting n best samples according to the cost calculated, (iv) update the covariance matrix $\Sigma^{\mathbf{d}}$ and the mean $\boldsymbol{\mu}^{\mathbf{d}}$ of the distribution based on the best n samples chosen. $\boldsymbol{\mu}^{\mathbf{d}}$ and $\Sigma^{\mathbf{d}}$ are updated such that the expected evaluation value decrease at every iteration. For further details on CMA-ES and its use, refer to [39].

IV. DATA-EFFICIENT VARIABLE IMPEDANCE LEARNING FRAMEWORK

The DEVILC framework utilizes GP models to learn the Cartesian impedance model of the system. The learned GP model is then used to optimize a NN-based impedance adaptation policy using CMA-ES. The cartesian impedance model represents the environment-robot dynamic relationship in (4). We learn the following Cartesian impedance model of the robot manipulator [40] using GP:

$$\mathbf{s}_{t+1} = \mathbf{s}_t + f(\mathbf{s}_t, \mathbf{u}_t) + \boldsymbol{\omega}\,. \tag{6}$$

where $\mathbf{s}_t$ is the state of the robot end-effector at time step t, $\mathbf{u}_t$ is the applied action, $\mathbf{s}_{t+1}$ is the next state, and $\boldsymbol{\omega}$ is the i.i.d Gaussian noise. The GP model with these inputs predicts the mean $\boldsymbol{\mu}(\mathbf{s}_{t+1})$ and variance $\boldsymbol{\sigma}^2(\mathbf{s}_{t+1})$ based on the current state and the action. We define the state $\mathbf{s}_t$ as $[\mathbf{x}_t, \dot{\mathbf{x}}_t]$, and action $\mathbf{u}_t$ as $[\mathbf{f}^t{}_{\mathbf{ext}}, \mathbf{K}_t]$. $\mathbf{K}_t$ is sampled from a parameterized impedance adaptation policy with $\boldsymbol{\pi}$ s.t $\mathbf{K}_t = \boldsymbol{\pi}\left(\mathbf{s}_t, \mathbf{f}^t{}_{\mathbf{ext}} \mid \boldsymbol{\theta}\right)$ which can be a NN. $\mathbf{f}^t{}_{\mathbf{ext}}$ is the sensed external force acting on the robot at time instant t, this is an uncertain external factor the VILC needs to compensate for. The damping parameters are chosen according to the critical damping condition, $\mathbf{D} = 2\sqrt{\mathbf{K}}$. For an N Degree of Freedom (DoF) Cartesian impedance dynamics considered, f contains N independent GP model with each GP model approximating the dynamics along one DoF.

We aim to optimize the compliant behavior of the robot end-effector can be optimized by designing a suitable variable impedance control strategy. Within the proposed DEVILC framework, given a manipulation task objective, this impedance adaptation strategy for the underlying VIC is optimized using CMA-ES and the GP-based Cartesian impedance model (6). We learn a NN based impedance adaptation policy $\boldsymbol{\pi}$ in an episodic MBRL setting, where after each episode of interaction with the real system, the GP model is updated and a policy is optimized using CMA-ES for the entire task horizon. The objective of compliant robot manipulation is defined to achieve manipulation task requirements/goals while executing a high level of compliance. In this work, we consider the scenario where the manipulation task requirement can be represented as tracking a desired robot state, but it is generalizable to any objective that can be measured. The compliance objective is to minimize the stiffness of the underlying VIC controller. A cost function describing the task objective and the compliance objective is designed for the real system,

$$C\left(\mathbf{s}_t, \mathbf{K}_t\right) = \delta\mathbf{s}_t^T \mathbf{Q}_t \delta\mathbf{s}_t + \lambda(\mathbf{K}_t)^T \mathbf{R}_t \lambda(\mathbf{K}_t)\,, \tag{7}$$

where $\mathbf{s}_t^r$ denotes the reference or goal states. $\lambda(\mathbf{K}_t)$ is the Eigenvalues of the stiffness matrix represented in a vector form, $\delta\mathbf{s}_t = \mathbf{s}_t^r - \mathbf{s}_t$ and $\mathbf{Q}_t$ and $\mathbf{R}_t$ are diagonal gain matrices for task and compliance components respectively. These gain matrices can be either constant or can be a function of the robot's states. For example, in the case of a reference tracking task $\mathbf{Q}_t$ can be chosen as a linear function of $\|\delta\mathbf{s}_t\|$ in order to have higher penalties for larger deviations from the target. While this cost is defined over the real system, it can be used as a target to train the reward function $r = -C$, i.e by taking the negative of this cost at every time instant.

Given the GP model f (6) and the cost function (7), the goal is to find the optimal policy parameters $\boldsymbol{\theta}$ that maximizes the reward over the entire task horizon given by:

$$J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{t=1}^{T} r\left(\mathbf{s}_t, \mathbf{K}_t\right) \mid \boldsymbol{\theta}\right]. \tag{8}$$

This is achieved by predicting the state evolution over the GP dynamics model. State-to-next-state propagation is carried out as in Monte Carlo estimation by sampling according to the GP model. But additionally, each of these rollouts is considered as a measurement of a function $G(\boldsymbol{\theta})$ that is the actual function $J(\boldsymbol{\theta})$ perturbed by a noise $N(\boldsymbol{\theta})$: [9]

$$\begin{aligned} G(\boldsymbol{\theta}) &= J(\boldsymbol{\theta}) + N(\boldsymbol{\theta}) \\ &= \sum_{t=1}^{T} r\left(f\left(\mathbf{s}_{t-1}, \mathbf{u}_{t-1}\right)\right)\,, \end{aligned} \tag{9}$$

here $\mathbf{K}_{t-1} = \pi\left(\mathbf{s}_{t-1}, \mathbf{f}^t_{\mathbf{ext}} \mid \boldsymbol{\theta}\right)$ and $\mathbf{f}^t_{\mathbf{ext}}$ is chosen from a random episode of interaction data. We would like to maximize its expectation:

$$\begin{aligned} \mathbb{E}[G(\boldsymbol{\theta})] &= \mathbb{E}[J(\boldsymbol{\theta}) + N(\boldsymbol{\theta})] \\ &= \mathbb{E}[J(\boldsymbol{\theta})] + \mathbb{E}[N(\boldsymbol{\theta})] \\ &= J(\boldsymbol{\theta}) + \mathbb{E}[N(\boldsymbol{\theta})]\,. \end{aligned} \tag{10}$$

With the assumption that $\mathbb{E}[N(\boldsymbol{\theta})] = 0$ thereby maximizing $\mathbb{E}[G(\boldsymbol{\theta})]$ is equivalent to maximizing $J(\boldsymbol{\theta})$. In order to search for the optimal policy $\boldsymbol{\pi}^{\star}$ with parameters $\boldsymbol{\theta}^{\star}$, we utilize the evolutionary optimization strategy, CMA-ES. CMA-ES solves the maximization of an objective function $J(\boldsymbol{\theta})$ as the optimization of the noisy function $G(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + N(\boldsymbol{\theta})$ where $N(\boldsymbol{\theta})$ is the noise [9], [11]. This facilitates maximizing the objective without computing or estimating it explicitly. CMA-ES performs four steps at each generation i:

1) sample β new candidates according to a multivariate Gaussian distribution of mean $\boldsymbol{\mu}^d{}_i$ and covariance $\boldsymbol{\Sigma}^d{}_i$.
2) rank the β sampled candidates based on their noisy performance over the objective $G(\boldsymbol{\theta}_i)$.
3) compute $\boldsymbol{\mu}^d{}_{i+1}$ by averaging n best candidates: $\boldsymbol{\mu}^d{}_{i+1} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\theta}_i$.
4) updates the covariance matrix to match the distribution of the $\boldsymbol{\mu}^d$ best candidates identified.

One key advantage of this ranking-based approach is it reduces the impact of noise on the performance function. This is because the solution is looking for the β best candidates and errors can only happen at the boundaries between the low-performing and high-performing solutions. Even if a candidate is misclassified because of the noise, this error is smoothed out by taking the average in step 3 to calculate $\boldsymbol{\mu}^d{}_{i+1}$ [41].

By combining the GP-based model learning, policy evaluation based on the noisy function (9) (10), and CMA-ES based policy search forms a MBRL based VILC framework DEVILC as shown in Fig. 2. The DEVILC method alternates episodically between the robot interacting with the real system and learning the impedance adaptation policy (which also includes updating the Cartesian impedance model). The DEVILC Algorithm is shown in Algorithm 1.

Algorithm 1 DEVILC

```
Given a cost function $C$, initialise an NN policy $\boldsymbol{\pi}$ .
Populate dataset $\mathcal{D}$ using a VIC with random $\mathbf{K}$ values
  for $N_i$ initial trials .
while task is not solved do
    Learn the GP dynamics model $f$ on $\mathcal{D}$ .;
    $\boldsymbol{\theta}^* = \mathrm{argmax}_{\boldsymbol{\theta}}\, \mathbb{E}[G(\boldsymbol{\theta})]$ (Section IV) .
    for $i \leftarrow 1$ to Task Horizon do
        $\mathbf{K}_i = \boldsymbol{\pi}\left(\mathbf{s}_i, \mathbf{f}^i{}_{\mathbf{ext}} \mid \boldsymbol{\theta}^\star\right)$ .
        $\mathbf{s}_{i+1}, \mathbf{f}^{i+1}{}_{\mathbf{ext}}$ = execute VIC with $\mathbf{K}_i$ ((5)) .
        $\mathcal{D} = \mathcal{D} \cup \left\{\mathbf{s}_{\mathbf{i}\mathcal{C}\mathbf{1}}, \mathbf{f}^{i+1}{}_{\mathbf{ext}}, \mathbf{K}_\mathbf{i}\right\}$
    end for
end while
```

V. EVALUATION

We evaluated the proposed DEVILC framework using a couple of simulation setups in addition to a real experiment. The *objective* of these experiments was to evaluate the effectiveness of the proposed approach to learning suitable VIC policy for robotic manipulation tasks. Four tasks of varying complexity chosen for evaluation are shown in Fig. 3. The experiments were set up such that, depending on the task, we consider the adaptation of the stiffness along the specific DoF of the robot manipulator while keeping the stiffness values along the other DoFs constant. This simplification of the experimental set-up will allow us to clearly explain the stiffness adaptation behaviour. The *criteria* for measuring the performance of DEVILC is based on how well the robot is able to perform the task, while being maximally compliant, given by (7).

GP models are used to learn the system dynamics, while a NN is used as the policy. The input state space for the GP model contains the Cartesian position and velocity. The input action space for the GP models contains the external forces acting on the end-effector ($\mathbf{f}_{\mathbf{ext}}$) and the stiffness values ($\mathbf{K}$). Given these inputs GP model predicts the next Cartesian pose ($\mathbf{x}$) and velocity ($\dot{\mathbf{x}}$) of the robot end-effector. The input state space of the NN policy contains the $\mathbf{x}$, $\dot{\mathbf{x}}$, and $\mathbf{f}_{\mathbf{ext}}$. The output of this NN policy is the predicted stiffness values. Dimensions of all these state and action spaces for the GP model and NN policy are dependent on the number of DoFs considered for the task. We use a GP structure with an exponential kernel with automatic relevance determination [40]. The NN policy in all the experiments contained one hidden layer with 32 neurons. For the CMA-ES optimizer we utilize BIPOP-CMA-ES with restarts [42] in combination with UH-CMA-ES for noisy functions [43] as proposed in [9].

The values for the damping component are chosen as $\mathbf{D} = 2\sqrt{\mathbf{K}}$. he mass matrix $\mathbf{M}$ is kept constant to avoid stability issues during the experiment. The sampling frequency which is equivalent to the VIC frequency is set as 10 Hz for all the tasks. For all the experiments 10 learning trials/episodes were conducted alternating between model learning and policy optimization. For the real-world experiments, the interaction data was downsampled to 20 data points per episode due to the low computational speed of the GP inference. For all chosen tasks, the requirements are defined for achieving a desired goal pose for the robot end-effector. However, the robot is also required to be highly compliant whenever it is possible or be stiff only when it is necessary. This is achieved by using a weighted reward in (7) for the task requirement (first term) and maximizing compliance (second term). In the considered tasks, we are, essentially, trying for a trade-off between position control and compliance. The cost function for each task has different values of the gain matrices $\mathbf{Q}_t$ and $\mathbf{R}_t$ in (7), defining the desired trade-off between position accuracy and compliant behavior. The values of these gain matrices are hand-tuned for each task and kept constant during the learning process. In all the results force is represented in Newtons and distance in cm.

A. SIMULATIONS

Two simulation experiments were conducted to evaluate the effectiveness of the proposed approach on learning VILC. We chose two manipulation tasks with different dynamics, (i) catching falling objects and (ii) pushing an object along a surface. In the first task, the robot has to adapt its impedance to optimally react to the impact of the falling object and also

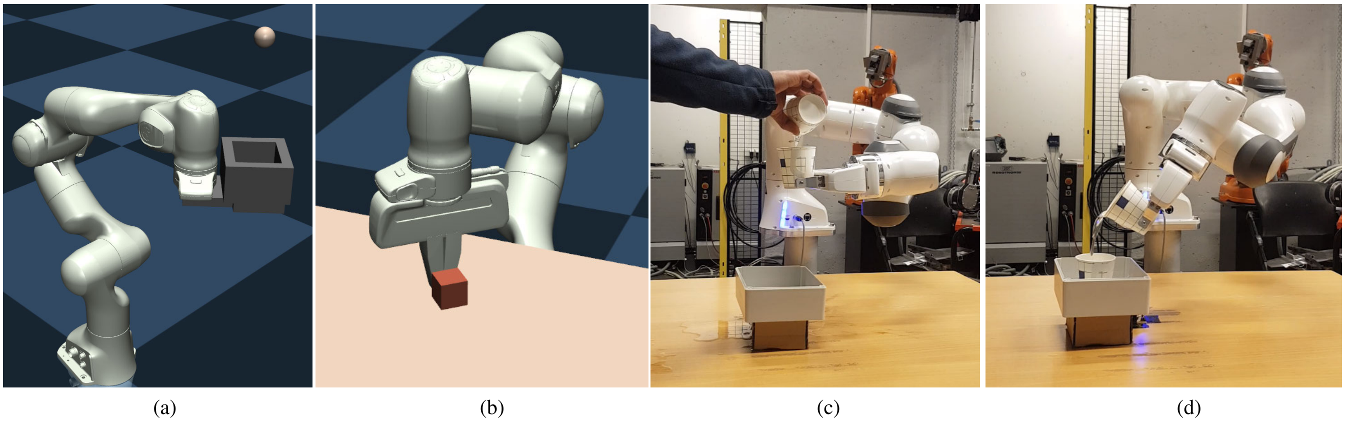


FIGURE 3. Simulation tasks, (a) Reacting to falling object: The robot manipulator with cup end-effector should hold a Cartesian position while smoothly catching a ball of weight $0.5\,\mathrm{g}$ falling into the cup. (b) Pushing task: A robot manipulator with a gripper end-effector should push an object over a rigid surface with friction to a target position. The two experimental tasks, (c) water filling: robot end-effector is fitted with an empty cup to which water is slowly filled by a person, (d) water pouring: robot end-effector is fitted with a cup filled with water and the robot transfers this water completely to another cup without any spillage.

to carry the additional weight added by the object. Whereas in the pushing task the robot has to adapt the impedance necessary to overcome the inertia of the object and the frictional forces and be more compliant towards the end of the task.

1) REACTING TO FALLING OBJECT

As shown in Fig. 3(a), the robot manipulator is fitted with a tray as the end-effector and an object is dropped to the tray first and then removed from the tray after one second. The robot is initialized to be highly compliant at rest position. Here the task requirement for the robot is to maintain its pose when the object is dropped to and removed from the tray while being maximally compliant as in (7). Multiple trials were performed with the object being dropped from different heights to the cup, resulting in robot behavior as shown in Fig. 4(a). The policy is optimized such that the deviation of the robot from its initial position is minimal while being as compliant as possible in reacting to the falling object. We only consider stiffness adaptation along the z direction for this task and the stiffness values along all other DoFs are kept unchanged during the learning. The result shows that the robot is at rest with low stiffness and the stiffness K_z is increased instantaneously in response to the impact force $\mathbf{f_{ext}}$ and further increases in response to deviation from the initial position. Upon removing the object, the stiffness is decreased enough to drive the robot back to the initial position. The maximum deviation of the robot from the desired pose upon impact is $0.1\,\mathrm{cm}$, before recovering back to rest pose at $t = 2\,\mathrm{s}$.

2) PUSHING TASK

The objective of this task to push an object placed on a table with friction to a target position of $10\,\mathrm{cm}$ as shown in Fig. 3(b). The policy is learned to adapt the stiffness in the pushing directions (x and y) to push the object to the target while stiff only when necessary and being compliant otherwise. The stiffness along other DoFs are kept constant. The robot is initialized with low stiffness values along the pushing directions. The results in Fig. 4(b) show that the robot pushes the object to the target pose in $1.4\,\mathrm{s}$. The stiffness is initially increased to larger values as expected along the pushing directions to overcome the inertia of the object. The stiffness is decreased when the object is close to the target position, executing high compliance. The robot learned to adapt the stiffness profiles in a suitable way to push the object to the target with high accuracy while being stiff only when necessary.

B. REAL-WORLD EXPERIMENTS

Two experiments were conducted to evaluate the effectiveness of the proposed approach in real-world robotic tasks demanding impedance adaptation. We chose the water pouring and filling task with the Franka-Emika Panda Robot manipulator. This is inspired by human manipulation behavior, we adapt our arm stiffness continuously in both pouring and fillings tasks to be efficient. For both filling and poring tasks, the cost function is described as reaching a target pose while being as compliant as in(7). For both tasks, we consider only stiffness adaptation along the z axis while maintaining a constant stiffness along all other DoFs.

1) WATER FILLING

The experimental setup is shown in Fig. 3(c) where the robot end-effector is fitted with an empty cup and water is poured into the cup. The robot is initialized with a low stiffness value to be highly compliant at the initial position. Here the task objective (7) for the robot is to maintain its pose by continuously increasing the stiffness, such that it is enough to hold the extra weights of the water being added to the cup. Results in Fig. 4(c) show that the learned VILC continuously increases the robot stiffness while water is added to the cup

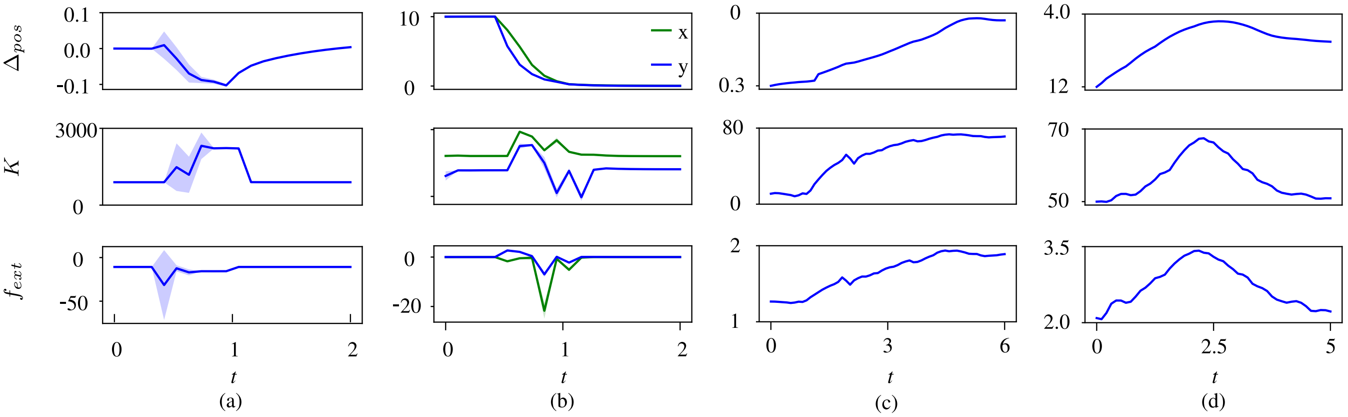


FIGURE 4. Evaluation results: (a) Reacting to falling object: the robot is expected to have minimal deviation Δ_{pos} from the initial pose in z direction. The results shown here are the mean values over 10 trials where the objects is dropped from randomly chosen heights between $(0.5 - 1.0)$ m. (b) Pushing task: the robot is tasked to push an object on the table to 10 cm in x and y directions at $t = 0.5$ s. The results shown here are over 10 trials with objects of random weights between $(0.5 - 1.0)$. (c) Water filing: water is filled into a cup fitted to the robot end-effector and the robot is expected to have a minimum deviation from the initial pose while executing compliance. The stiffness is varied only z direction and is kept constant in along all other DoF, all the values shown here are along z direction. (d) Water pouring: the robot is tasked to pour the water into a cup placed on the table. The pouring task is defined by commanding the robot to move to a pre-defined goal pose. The stiffness is varied only in z direction and is kept constant in along all other DoF, all the values shown here are along z direction.

allowing only a small deviation of only 3 mm from the desired pose. The stiffness profile generated by the VILC during the tasks appears to be very well correlated with the sensed external force on the end-effector imparted by the water.

2) WATER POURING

The experimental setup for the task is shown in Fig. 3(d) where the robot end-effector is fitted with a cup filled with water. The VIC is initialized with the right amount of compliance such that the robot holds the cup filled with water at the initial pose. The pouring task is defined as the robot pouring the water into a second cup placed on the table. The robot movement for the pouring task is defined as reaching a target end-effector pose such that the water is entirely transferred to the second cup. The task objective (7) here is to reach the target end-effector pose of 8 cm while being as compliant as possible. The robot is expected to learn to be less stiff as the water is transferred to the second cup. Results in Fig. 4(d) show that the learned VILC increases the robot stiffness in relation to the increased sensed forces during the first phase of the task where the cup is tilted to start transferring the water to the second cup. In the second phase of the task, (i.e once the water starts to flow into the second cup), the stiffness is decreased in relation to the decreased weight of the water robot has to hold. Overall the learned VILC policy is able to reach the target pose while being compliant.

VI. DISCUSSION

The VILC approach presented in the work is evaluated on different tasks in Section V to learn impedance adaptation strategies. The optimization objective in all experiments has been to maintain a high level of compliance in general while being stiff only when demanded by the task. In all the tasks, the task requirement is defined by achieving a desired goal pose for the robot end-effector. The performance of the impedance adaptation strategy is evaluated based on how well it is able to achieve this requirement while being maximally compliant. In the case of tasks demanding positional accuracy, this means a suitable trade-off between accuracy and compliance. But IC under stable behavior allows the robot to asymptotically converge to the target pose. This property allows the learning methods to suitably vary the impedance to be maximally compliant without necessarily sacrificing the positional accuracy, especially in tasks that do not demand strict real-time trajectory tracking. Whereas optimizing impedance profiles to be maximally compliant allows robots to be more dexterous, safe, and energy efficient. The NN policy obtained using CMA-ES based optimization could adapt the stiffness profile in response to external forces and deviations from the target

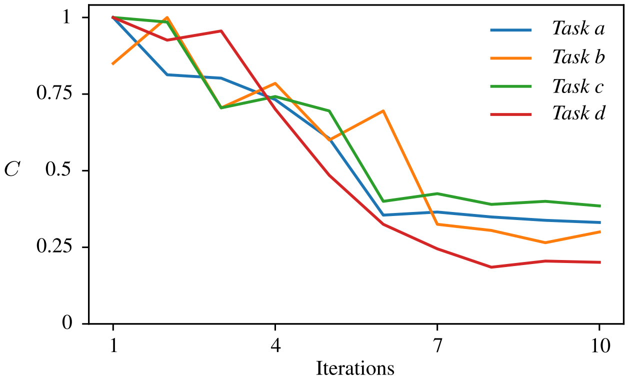


FIGURE 5. DEVILC training results for all the tasks in Fig. 3. Mean cumulative cost/ negative of the reward (i.e, $C = -r$) after each learning iteration for all tasks (a) - (d) in Fig. 3. The values are normalized between [0,1] for each task.

pose while maintaining a high level of compliance whenever possible.

Compared to the existing approach, the main advantage of our VILC approach is data efficiency as the stiffness adaptation policy is learned from a handful of trials without any constraint on the optimization objective and policy structure. The existing VILC methods with comparable data-efficiency are only PILCO based approaches and PI^2 [28]. While the PILCO-based approaches in [10] and [33] offers a highly data-efficient approach is limited by the type of cost functions and a differentiable policy and higher computational effort on optimizing the policy. Whereas, the proposed approach is generalizable to any policy and cost structure. Whereas PI^2 approach in [28] is not directly applicable to the force-based VIC considered in this work. CMA-ES based policy optimization is shown to achieve similar performance to PILCO and PI^2 in robotic manipulation tasks while being more data-efficient [9]. Because of these reasons, we have not provided any comparison with these approaches in this work. Although there are other RL approaches using complex dynamical models such as NN as discussed in Section II, they pose challenges to real-world applications due to low data efficiency. The training results for all tasks in simulation and experiments shown in Fig. 5 demonstrates the data-efficiency of the DEVILC framework. For all four tasks, the DEVILC framework optimizes an impedance adaptation policy within 10 iterations, which corresponds to 200 data samples in total. The NN policy obtained is computationally efficient to evaluate in real-time, even with larger NN structures, thanks to its ability to leverage efficient parallel computation using GPUs or TPUs.

We have not discussed the aspects of safety/stability in this work. We assumed a stiffness parameter range learned is with the stable region for the underlying VIC. Guaranteeing stability properties to the resulting VILC is challenging as guarantees have to be provided in real-time in an online fashion as the stiffness values predicted by the policy are state-dependent. The approach proposed in [44] by designing a quadratic Lyapunov candidate function could be coupled with GP models to provide probabilistic stability guarantees similar to safety guarantees in [45]. GP models allow for providing such guarantees on safety and stability by using additional optimization constraints [46]. But this needs further research in the case of VILC for providing closed-loop safety and stability guarantees during learning and for the final policy. The safe learning approaches described in [46] are interesting to explore for model-based VILC. One feasible approach in this direction could be to provide probabilistic safety guarantees using Control Lyapunov Functions (CLF) for stability and Control Barrier Functions (CBF) as a safety filter to solve constrained optimization problems over the GP model [45].

The strength of the proposed approach relies on a trade-off between high data efficiency and scalability to complex problems demanding richer model representations. GP models have high data efficiency [8], providing a reliable estimate of model uncertainties which is very suitable for model-based policy optimization [8] and MBRL in general. GP have limited ability in modeling complex dynamics and is not well suited for highly noisy data. Additionally, GP models are not good at representing non-smooth dynamics such as contact dynamics or human-robot-interaction dynamics. This effect was observed in task (b) where the robot is pushing an object to a goal position. Here the robot end effector is prone to lose contact with the object during the task and making it difficult for the GP model to learn the dynamics. This results in slightly noisy impedance profiles in Fig. 4(b). GP models also poses challenges in computational effort when the dataset is large. Similarly, CMA-ES limits the size of the policy parameter size for computational speed, which is a common drawback of most evolutionary algorithms. More scalable model-based VILC approaches can be developed using Deep Neural Networks (DNN) models and RL. Still, they have much higher sample complexity and low scope of providing safety guarantees.

VII. CONCLUSION

In this work, we presented DEVILC, a data-efficient model-based VILC approach to learning compliant robotic manipulations skills. The Cartesian impedance dynamics of the robot controlled using a VIC is learned using GP models. The learned dynamics model was coupled with CMA-ES optimization strategy to find a suitable impedance adaptation policy for a task. The optimization objective was designed such that the robot should be compliant unless it is necessary to be stiff, which is fundamental to how humans manipulate objects. We evaluated our approach to simplified robotic manipulation tasks in simulations and experiments. The impedance adaptation policy optimized exhibited the desired compliance behavior by being highly compliant unless acted upon by an external force or the robot pose deviated from the desired pose. In future work, we aim to extend this approach to incorporate safety and stability constraints.

REFERENCES

[1] E. Bizzi, N. Accornero, W. Chapple, and N. Hogan, "Posture control and trajectory formation during arm movement," *J. Neurosci.*, vol. 4, no. 11, pp. 2738–2744, Nov. 1984.

[2] N. Hogan, "An organizing principle for a class of voluntary movements," *J. Neurosci.*, vol. 4, no. 11, pp. 2745–2754, Nov. 1984.

[3] S. D. Kennedy and A. B. Schwartz, "Stiffness as a control factor for object manipulation," *J. Neurophysiol.*, vol. 122, no. 2, pp. 707–720, Aug. 2019.

[4] N. Hogan, "Impedance control: An approach to manipulation," in *Proc. Amer. Control Conf.*, Jul. 1984, pp. 304–313.

[5] O. Khatib, "A unified approach for motion and force control of robot manipulators: The operational space formulation," *IEEE J. Robot. Autom.*, vol. RA-3, no. 1, pp. 43–53, Feb. 1987.

[6] R. Ikeura and H. Inooka, "Variable impedance control of a robot for cooperation with a human," in *Proc. IEEE Int. Conf. Robot. Automat.*, vol. 3, May 1995, pp. 3097–3102.

[7] F. J. Abu-Dakka and M. Saveriano, "Variable impedance control and learning—A review," *Frontiers Robot. AI*, vol. 7, Dec. 2020, Art. no. 590681.

[8] M. Deisenroth and C. E. Rasmussen, "PILCO: A model-based and data-efficient approach to policy search," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 465–472.

[9] K. Chatzilygeroudis, R. Rama, R. Kaushik, D. Goepp, V. Vassiliades, and J.-B. Mouret, "Black-box data-efficient policy search for robotics," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 51–58.

[10] A. S. Anand, M. Hagen Myrestrand, and J. T. Gravdahl, "Evaluation of variable impedance- and hybrid force/motioncontrollers for learning force tracking skills," in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Jan. 2022, pp. 83–89.

[11] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evol. Comput.*, vol. 9, no. 2, pp. 159–195, Jun. 2001.

[12] S. Calinon, I. Sardellitti, and D. G. Caldwell, "Learning-based control strategy for safe human–robot interaction exploiting task and robot redundancies," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2010, pp. 249–254.

[13] S. M. Khansari-Zadeh, K. Kronander, and A. Billard, "Modeling robot discrete movements with state-varying stiffness and damping: A framework for integrated motion generation and impedance control," in *Proc. Robot., Sci. Syst. X (RSS)*, vol. 10, 2014, p. 2014.

[14] D. Lee and C. Ott, "Incremental kinesthetic teaching of motion primitives using the motion refinement tube," *Auto. Robots*, vol. 31, nos. 2–3, pp. 115–131, Oct. 2011.

[15] M. Saveriano, S. An, and D. Lee, "Incremental kinesthetic teaching of end-effector and null-space motion primitives," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 3570–3575.

[16] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–35, Mar. 2018.

[17] P. Kormushev, S. Calinon, and D. Caldwell, "Reinforcement learning in robotics: Applications and real-world challenges," *Robotics*, vol. 2, no. 3, pp. 122–148, Jul. 2013.

[18] C.-C. Cheah and D. Wang, "Learning impedance control for robotic manipulators," *IEEE Trans. Robot. Autom.*, vol. 14, no. 3, pp. 452–465, Jun. 1998.

[19] A. Gams, B. Nemec, A. J. Ijspeert, and A. Ude, "Coupling movement primitives: Interaction with the environment and bimanual tasks," *IEEE Trans. Robot.*, vol. 30, no. 4, pp. 816–830, Aug. 2014.

[20] F. J. Abu-Dakka, B. Nemec, J. A. Jørgensen, T. R. Savarimuthu, N. Krüger, and A. Ude, "Adaptation of manipulation skills in physical contact with the environment to reference force profiles," *Auto. Robots*, vol. 39, no. 2, pp. 199–217, Aug. 2015.

[21] A. Kramberger, E. Shahriari, A. Gams, B. Nemec, A. Ude, and S. Haddadin, "Passivity based iterative learning of admittance-coupled dynamic movement primitives for interaction with changing environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 6023–6028.

[22] Y. Zhang, B. Chu, and Z. Shu, "A preliminary study on the relationship between iterative learning control and reinforcement learning," *IFAC-PapersOnLine*, vol. 52, no. 29, pp. 314–319, 2019.

[23] R. Martín-Martín, M. A. Lee, R. Gardner, S. Savarese, J. Bohg, and A. Garg, "Variable impedance control in end-effector space: An action space for reinforcement learning in contact-rich tasks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 1010–1017.

[24] C. C. Beltran-Hernandez, D. Petit, I. G. Ramirez-Alpizar, and K. Harada, "Variable compliance control for robotic peg-in-hole assembly: A deep-reinforcement-learning approach," *Appl. Sci.*, vol. 10, no. 19, p. 6923, Oct. 2020.

[25] C. C. Beltran-Hernandez, D. Petit, I. G. Ramirez-Alpizar, T. Nishi, S. Kikuchi, T. Matsubara, and K. Harada, "Learning force control for contact-rich manipulation tasks with rigid position-controlled robots," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5709–5716, Oct. 2020.

[26] M. Bogdanovic, M. Khadiv, and L. Righetti, "Learning variable impedance control for contact sensitive tasks," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 6129–6136, Oct. 2020.

[27] P. Varin, L. Grossman, and S. Kuindersma, "A comparison of action spaces for learning manipulation tasks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 6015–6021.

[28] J. Buchli, F. Stulp, E. Theodorou, and S. Schaal, "Learning variable impedance control," *Int. J. Robot. Res.*, vol. 30, no. 7, pp. 820–833, 2011.

[29] M. Saveriano, F. J. Abu-Dakka, A. Kramberger, and L. Peternel, "Dynamic movement primitives in robotics: A tutorial survey," 2021, *arXiv:2102.03861*.

[30] E. Theodorou, J. Buchli, and S. Schaal, "Learning policy improvements with path integrals," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 828–835.

[31] M. Kalakrishnan, L. Righetti, P. Pastor, and S. Schaal, "Learning force control policies for compliant manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 4639–4644.

[32] M. Kim, S. Niekum, and A. D. Deshpande, "SCAPE: Learning stiffness control from augmented position control experiences," in *Proc. Conf. Robot Learn.*, 2022, pp. 1512–1521.

[33] C. Li, Z. Zhang, G. Xia, X. Xie, and Q. Zhu, "Efficient force control learning system for industrial robots based on variable impedance control," *Sensors*, vol. 18, no. 8, p. 2539, Aug. 2018.

[34] M. H. Myrestrand, "Learning compliant robotic manipulation," in *Institutt for Teknisk Kybernetikk [2862], NTNU Open*. Trondheim, Norway: NTNU, 2021.

[35] L. Roveda, J. Maskani, P. Franceschi, A. Abdi, F. Braghin, L. M. Tosatti, and N. Pedrocchi, "Model-based reinforcement learning variable impedance control for human–robot collaboration," *J. Intell. Robotic Syst.*, vol. 100, no. 2, pp. 417–433, Nov. 2020.

[36] A. S. Anand, J. T. Gravdahl, and F. J. Abu-Dakka, "Model-based variable impedance learning control for robotic manipulation," *Robot. Auto. Syst.*, vol. 170, Dec. 2023, Art. no. 104531.

[37] H.-P. Huang and S.-S. Chen, "Compliant motion control of robots by using variable impedance," *Int. J. Adv. Manuf. Technol.*, vol. 7, no. 6, pp. 322–332, Dec. 1992.

[38] L. Villani and J. De Schutter, "Force control," in *Springer Handbook of Robotics*. Berlin, Germany: Springer, 2016, pp. 195–220.

[39] N. Hansen, "The CMA evolution strategy: A tutorial," 2016, *arXiv:1604.00772*.

[40] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*, vol. 2, no. 3. Cambridge, MA, USA: MIT Press, 2006.

[41] Y. Jin and J. Branke, "Evolutionary optimization in uncertain environments—A survey," *IEEE Trans. Evol. Comput.*, vol. 9, no. 3, pp. 303–317, Jun. 2005.

[42] N. Hansen, "Benchmarking a BI-population CMA-ES on the BBOB-2009 function testbed," in *Proc. 11th Annu. Conf. Companion Genetic Evol. Comput. Conf., Late Breaking Papers*, Jul. 2009, pp. 2389–2396.

[43] N. Hansen, A. S. P. Niederberger, L. Guzzella, and P. Koumoutsakos, "A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion," *IEEE Trans. Evol. Comput.*, vol. 13, no. 1, pp. 180–197, Feb. 2009.

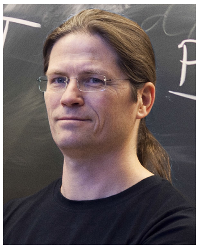
[44] Z. Jin, A. Liu, W. Zhang, and L. Yu, "An optimal variable impedance control with consideration of the stability," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1737–1744, Apr. 2022.

[45] M. J. Khojasteh, V. Dhiman, M. Franceschetti, and N. Atanasov, "Probabilistic safety constraints for learned high relative degree system dynamics," in *Proc. 2nd Conf. Learn. Dyn. Control*, 2020, pp. 781–792.

[46] A. Anand, K. Seel, V. Gjærum, A. Håkansson, H. Robinson, and A. Saad, "Safe learning for control using control Lyapunov functions and control barrier functions: A review," *Proc. Comput. Sci.*, vol. 192, pp. 3987–3997, Oct. 2021.

AKHIL S. ANAND (Member, IEEE) received the B.Tech. degree in mechanical engineering from the Indian Institute of Technology (IIT), Patna, India, in 2013, the M.Sc. degree in mechatronics from the University of Siegen, Germany, in 2018, and the Ph.D. degree from the Norwegian University of Science and Technology (NTNU), in 2023. He is currently a Postdoctoral Researcher with the Department of Engineering Cybernetics, NTNU. His research interests include robotics, reinforcement learning, data-driven control, and optimization.

RITURAJ KAUSHIK received the B.Tech. degree in electronics and communication engineering and the M.Tech. degree in electronics design and technology from Tezpur University, India, in 2012 and 2016 respectively, and the Ph.D. degree from Université de Lorraine, France, in 2020. He is currently a Postdoctoral Researcher with the Intelligent Robotics Group, Aalto University, Finland. His research interests include data-efficient robot learning, evolutionary robotics, sim-to-real robot learning, and reinforcement learning.



JAN TOMMY GRAVDAHL (Senior Member, IEEE) was born in 1969. He received the Siv.ing. and Dr.ing. degrees in engineering cybernetics from the Norwegian University of Science and Technology (NTNU), in 1994 and 1998, respectively. From 1998 to 2001, he was a Postdoctoral Researcher with the Department of Engineering Cybernetics, NTNU, where he was appointed as an Associate Professor, in 2001, a Professor, in 2005, the Deputy Department Head, from 2006 to 2007, and the Department Head, from 2008 to 2009. From 2007 to 2008, he was with the Centre for Complex Dynamic Systems and Control (CDSC), The University of Newcastle, Australia. He is currently a Professor with NTNU. His current research interests include mathematical modeling and nonlinear control in general and with the application to turbomachinery, spacecraft, robots, ships, and nanopositioning devices. He was a recipient of the IEEE Transactions on Control Systems Technology Outstanding Paper Award, in 2000 and 2017.

FARES J. ABU-DAKKA (Member, IEEE) received the B.Sc. degree in mechanical engineering from Birzeit University, Palestine, in 2003, and the D.E.A. and Ph.D. degrees in robotics motion planning from the Polytechnic University of Valencia, Spain, in 2006 and 2011, respectively.

In 2012, he was a Postdoctoral Researcher with the Jozef Stefan Institute, Slovenia. From 2013 to 2016, he was a Visiting Professor with ISA, Carlos III University of Madrid, Spain. From 2016 to 2019, he was a Postdoctoral Researcher with the Istituto Italiano di Tecnologia (IIT). From 2019 to 2022, he was a Research Fellow with Aalto University. Then, in 2022, he moved to MIRMI, Technical University of Munich, Germany, as a Senior Scientist and a Leader of the Robot Learning Group. He is currently a Lecturer and a Researcher with the Faculty of Engineering, Mondragon University, Spain. His research interests include the intersection of control theory, differential geometry, and machine learning, in order to enhance robot manipulation performance and safety.

Dr. Abu-Dakka served as an Associate Editor for *ICRA*, *IROS*, and IEEE Robotics and Automation Letters. For more information visit the link: https://sites.google.com/view/abudakka/.