



Doctoral Thesis in Computer Science

Data-Driven Methods for Contact-Rich Manipulation: Control Stability and Data-Efficiency

SHAHBAZ ABDUL KHADER

Data-Driven Methods for Contact-Rich Manipulation: Control Stability and Data-Efficiency

SHAHBAZ ABDUL KHADER

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Friday the 17th September 2021, at 2:00 p.m. in F3, Lindstedsvägen 26, Stockholm.

Doctoral Thesis in Computer Science
KTH Royal Institute of Technology
Stockholm, Sweden 2021

© Shahbaz Abdul Khader
© Danica Kragic, Paper A,B,C,D
© Hang Yin, Paper A,B,C,D
© Pietro Falco, Paper A,B,C,D

ISBN 978-91-7873-937-0
TRITA-EECS-AVL-2021:49

Printed by: Universitetsservice US-AB, Sweden 2021

Abstract

Autonomous robots are expected to make a greater presence in the homes and workplaces of human beings. Unlike their industrial counterparts, autonomous robots have to deal with a great deal of uncertainty and lack of structure in their environment. A remarkable aspect of performing manipulation in such a scenario is the possibility of physical contact between the robot and the environment. Therefore, not unlike human manipulation, robotic manipulation has to manage contacts, both expected and unexpected, that are often characterized by complex interaction dynamics.

Skill learning has emerged as a promising approach for robots to acquire rich motion generation capabilities. In skill learning, data driven methods are used to learn reactive control policies that map states to actions. Such an approach is appealing because a sufficiently expressive policy can almost instantaneously generate appropriate control actions without the need for computationally expensive search operations. Although reinforcement learning (RL) is a natural framework for skill learning, its practical application is limited for a number of reasons. Arguably, the two main reasons are the lack of guaranteed control stability and poor data-efficiency. While control stability is necessary for ensuring safety and predictability, data-efficiency is required for achieving realistic training times. In this thesis, solutions are sought for these two issues in the context of contact-rich manipulation.

First, this thesis addresses the problem of control stability. Despite unknown interaction dynamics during contact, skill learning with stability guarantee is formulated as a model-free RL problem. The thesis proposes multiple solutions for parameterizing stability-aware policies. Some policy parameterizations are partly or almost wholly deep neural networks. This is followed by policy search solutions that preserve stability during random exploration, if required. In one case, a novel *evolution strategies*-based policy search method is introduced. It is shown, with the help of real robot experiments, that *Lyapunov stability* is both possible and beneficial for RL-based skill learning.

Second, this thesis addresses the issue of data-efficiency. Although data-efficiency is targeted by formulating skill learning as a model-based RL problem, only the model learning part is addressed. In addition to benefiting from the data-efficiency and uncertainty representation of the Gaussian process, this thesis further investigates the benefits of adopting the structure of *hybrid automata* for learning forward dynamics models. The method also includes an algorithm for predicting long-term trajectory distributions that can represent discontinuities and multiple modes. The proposed method is shown to be more data-efficient than some state-of-the-art methods.

Keywords: Skill Learning, Reinforcement Learning, Contact-Rich Manipulation

Sammanfattning

Autonoma robotar förväntas utgöra en allt större närvaro på människors arbetsplatser ioch i deras hem. Till skillnad från sina industriella motparter, behöver dessa autonoma robotar hantera en stor mängd osäkerhet och brist på struktur i sina omgivningar. En väsentlig del av att utföra manipulation i dylika scenarier, är förekomsten av fysisk interaktion med direkt kontakt mellan roboten och dess omgivning. Därför måste robotar, inte olik människor, kunna hantera både förväntade och oväntade kontakter med omgivningen, som ofta karaktäriseras av komplex interaktionsdynamik.

Skill learning, eller inläring av färdigheter, står ut som ett lovande alternativ för att låta robotar tillgodogöra sig en rik förmåga att generera rörelser. I *Skill Learning* används datadrivna metoder för att lära in en reaktiv *policy*, en reglerfunktion som kopplar tillstånd till styrsignaler. Detta tillvägagångssätt är tilltalande eftersom en tillräckligt uttrycksfull *policy* kan generera lämpliga styrsignaler nästan instantant, utan att behöva genomföra beräkningsmässigt kostsamma sökoperationer. Även om *Reinforcement Learning* (RL), förstärkningsinläring, är ett naturligt ramverk för *skill learning*, har dess praktiska tillämpningar varit begränsade av ett antal anledningar. Det kan med fog påstås att de två främsta anledningarna är brist på garanterad stabilitet, och dålig dataeffektivitet. Stabilitet i reglerloopen är nödvändigt för att kunna garantera säkerhet och förutsägbarhet, och dataeffektivitet behövs för att uppnå realistiska inläringstider. I denna avhandling söker vi efter lösningar till dessa problem i kontexten av manipulation med rik förekomst av kontakter.

Denna avhandling behandlar först problemet med stabilitet. Trots att dynamiken för interaktionen är okänd vid förekomsten av kontakter, formuleras *skill learning* med stabilitetsgarantier som ett modellfritt RL-problem. Avhandlingen presenterar flera lösningar för att parametrisera stabilitetsmedvetna *policies*. Detta följs sedan av lösningar för att söka efter *policies* som är stabila under slumpmässig sökning, om detta behövs. Några parametriseringar består helt eller delvis av djupa neurala nätverk. I ett fall introduceras också en sökmätod baserad på *Evolution Strategy*. Vi visar, genom experiment på faktiska robotar, att lyaponovstabilitet är både möjligt och fördelaktigt vid RL-baserad *skill learning*.

Vidare tar avhandlingen upp dataeffektivitet. Även om dataeffektiviteten angrips genom att formulera *skill learning* som ett modellbaserat RL-problem, så behandlar vi endast delen med modellinläring. Utöver att dra nytta av dataeffektiviteten och osäkerhetsrepresentationen i gaussiska processer, så undersöker avhandlingen även fördelarna med att använda strukturen hos hybrida automata för att lära in modeller för framåtdynamiken. Metoden innehåller även en algoritm för att förutsäga fördelningarna av trajektorier över en längre tidsrymd, för att representera diskontinuiteter och multipla moder. Vi visar att den föreslagna metodiken är mer dataeffektiv än ett antal existerande metoder.

For my wife Faiza

Acknowledgement

After spending several years in the industry, I decided to pursue graduate studies to challenge myself and develop further. This decision culminated in my PhD studies and what an enriching experience it was! It was an arduous journey though and I could not have done it without the support and guidance of so many wonderful people.

First and foremost, I wish to thank my main supervisor, Danica Kragic who supported me at every occasion. Thank you for admitting me to your group and giving me all the freedom that a PhD student can ask for. Your guidance and counselling enabled me to make the necessary transition from an engineer to a researcher. I also thank you for helping me improve my writing style. I cannot thank Hang, my co-supervisor, enough who made a crucial impact on my development as a PhD candidate. This was because you never hesitated to challenge me. Thank you for that! I am also indebted to Pietro, my other co-supervisor, whose consistent encouragement and wise suggestions proved extremely useful. I thank Hang and Pietro for the numerous discussions and constructive feedback that contributed towards a successful PhD.

This PhD would not have been possible without the generous support from the management of ABB Corporate Research, Sweden, where I was employed for the entirety of my studies. Not many companies would allow their employees to pursue studies on a full-time basis. Yet, thanks to Jonas and Xiaolong, that is exactly what happened. This is highly appreciated! I would also like to thank all my colleagues at ABB for making my life as a student comfortable and pleasant. I am most indebted to Ludde, my fellow PhD candidate at ABB, for all the cheerful experiences that we shared. I will have fond memories of our shared trips to WASP related events, including the study trips, especially the one to the silicon valley.

Being an industry-employed PhD student meant that I had to limit my time on the university campus. But, thanks to the wonderful people at RPL, I gained the most out of my time at KTH. Rika, Ali, Diogo, Ioanna, Mia, Isac and Yiannis, thank you for all the fruitful discussions we had.

I express my deep appreciation to the administrators of Wallenberg AI, Autonomous Systems and Software Program (WASP) for partially funding my PhD. A big thanks to my fellow WASP students of batch one of the AS track for making

my WASP community experience exciting and rewarding.

Finally, I would like to thank my wife for her endless support and understanding without which I would not have been able to succeed.

Contents

Acknowledgement	v
Contents	1
I Overview	3
1 Introduction	5
1.1 Motivation	5
1.2 Thesis Contributions	9
1.3 Thesis Outline	11
2 Background	13
2.1 Contact-Rich Manipulation	13
2.2 Classical Approaches for Contact-Rich Manipulation	15
2.3 Robot Skill Learning	16
2.4 Learning Contact-Rich Manipulation Skills	19
2.5 Control Stability in Skill Learning	21
2.6 Data-Efficiency in Skill Learning	24
3 Summary of Included Papers	29
4 Discussions and Conclusion	37
4.1 Control Stability	37
4.2 Data-Efficiency	39
4.3 Skill Learning	40
Bibliography	43

II Included Publications	53
A Stability-Guaranteed Reinforcement Learning for Contact-Rich Manipulation	A1
A.1 Introduction	A1
A.2 Related Works	A3
A.3 Background and Preliminaries	A5
A.4 Approach	A7
A.5 Experimental Results	A10
A.6 Discussions	A17
A.7 Conclusion	A18
B Learning Stable Normalizing-Flow Control for Robotic Manipulation	B1
B.1 Introduction	B1
B.2 Preliminaries	B3
B.3 Stable Normalizing-Flow Policy	B5
B.4 Experimental Results	B8
B.5 Discussions and Conclusion	B14
C Learning Deep Neural Policies with Stability Guarantees	C1
C.1 Introduction	C1
C.2 Related Work	C3
C.3 Preliminaries	C4
C.4 Methodology	C7
C.5 Experimental Results	C11
C.6 Discussions and Conclusion	C16
D Data-Efficient Model Learning and Prediction for Contact-Rich Manipulation Tasks	D1
D.1 Introduction	D1
D.2 Related Work	D2
D.3 Background and Problem Formulation	D4
D.4 Model Learning and Long-term Prediction	D6
D.5 Experimental Results	D12
D.6 Discussions	D17
D.7 Conclusion	D18

Part I

Overview

Chapter 1

Introduction

1.1 Motivation

The possibility of creating machines with human-like intelligence and skills has always intrigued humans. Today, it is not difficult to imagine the potential impact of such machines, or robots, on manufacturing, healthcare, transportation, exploration or even entertainment. To function effectively in a real-world environment, one which is characterized by uncertainty and lack of structure, a robot would need a high degree of autonomy. An *autonomous robot* has to perceive the world through a suit of sensors, build internal models of the external environment, plan according to the goal and resource constraints, and finally act through the set of available actuators. The more complex and dynamic the environment is, the more sophisticated the algorithms and software systems of the robot are. Unfortunately, despite decades of research, it is only systems at the lower end of autonomy that are most successful. For example, robots that work in highly structured environments, thus having little need for autonomy, are ubiquitous in the automobile manufacturing industry; whereas, service and social robots that require more autonomy to interact with uncertain and unstructured environments are relatively rare. It is well understood that only the successful application of *artificial intelligence (AI)*, the primary tool for achieving autonomy, can realize the long cherished dream of robots cohabiting with humans in homes and workplaces.

Although scientific research often targets specific areas of autonomous robots, it is beneficial to have an overall conceptual architecture of such a system in mind. After all, without such an architecture, no actual system can be built. To manage the complexity of an autonomous robot, a popular architecture that is often adopted is the *three tiers (3T)* architecture [1]. A simplified sketch is shown in Fig. 1.1. The *planning layer* is responsible for the generation of long-term plans required for a task. The *execution layer* breaks down a plan into a hierarchical structure consisting of behaviors which are then executed either concurrently or sequentially. The execution of behaviors is monitored and any

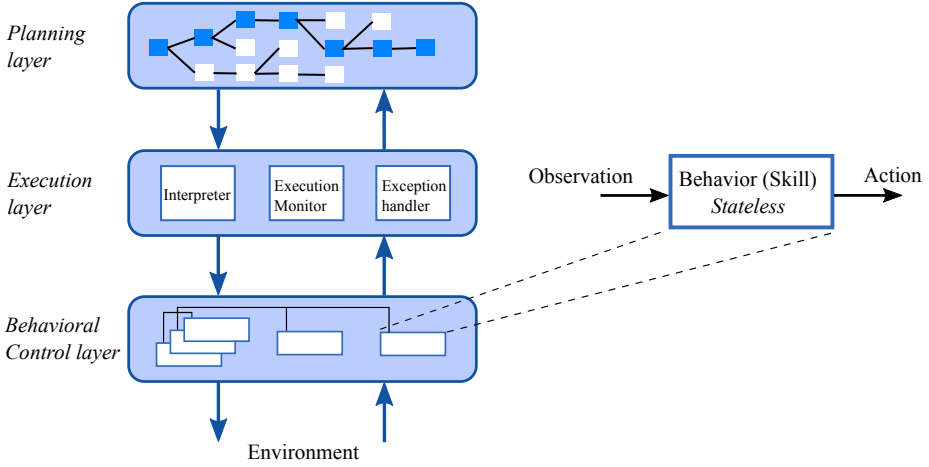
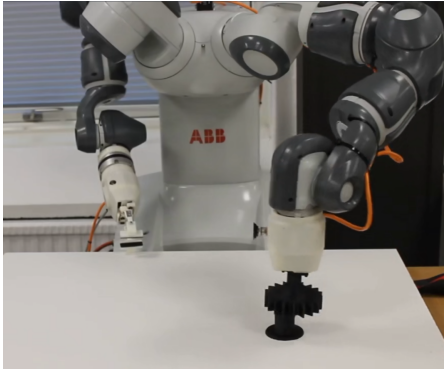


Figure 1.1: The three tier (3T) architecture for autonomous robots

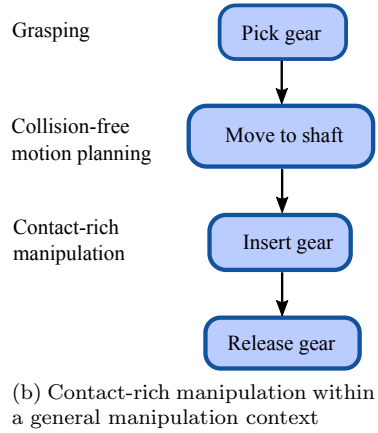
exception is also handled. The *behavioral control layer* is responsible for the execution of all behaviors that are activated by the execution layer. Behaviors are stateless perception-actuation loops running at high frequencies that usually do not perform any planning or search operations. It is here traditional control algorithms reside.

While the planning and the execution layers are the domain of AI planning methods, the behavioral control layer usually consists of hand-crafted perception-actuation control loops. Well-known algorithms such as the Kalman filter and the proportional-integral-derivative (PID) controller are good examples. However, there has been an increasing trend towards pushing more and more functionalities into the behavioral control layer. This can be motivated as follows. Consider a pick and place robotic manipulation task. In a simple design, the planning layer would synthesize the grasp [2] and motion path [3] solutions that satisfy the goal while avoiding obstacles. The main behaviors would be trajectory tracking control and gripper actuation. However, if there happens to be dynamic obstacles, it makes sense to endow the behavior layer with an online obstacle avoidance behavior. Then the planning layer could be reserved for higher level functions such as planning the order of multiple picks and places. The true potential of this strategy emerges when we consider leveraging the latest advances in deep learning to obtain a rich repertoire of behaviors, or *skills*, that are fast and reactive—unlike the traditional AI-based planning methods. Robot *skill learning*, also referred to as *robot learning* [4], thus has an important role to play in autonomous robots.

An unavoidable consequence of manipulating objects in an uncertain and unstructured environment is the possibility of making extensive contact with the environment. While traditional industrial robots mostly move in free space with high speeds, autonomous robots are expected to be able to control motion during



(a) A YuMi robot performing gear assembly



(b) Contact-rich manipulation within a general manipulation context

Figure 1.2: An example of contact-rich manipulation: gear assembly

contact. Motion of a robot in contact with its environment is called *constrained motion* and the manipulation process under such a condition is referred to as *contact-rich manipulation*. To understand contact-rich manipulation, consider the task of assembling a gear onto a fixed shaft as shown in Fig. 1.2. One can infer that in addition to grasping and collision free motion planning, the robot has to execute a compliant search operation to engage the gear onto the top part of the shaft and then proceed with a compliant insertion motion. Both the search and insert motions require *compliance control* due to uncertainty in the relative positions and orientations of the two mating pieces. Compliance control implies that the robot has to comply with motion constraints imposed by the environment in addition to generating motion in unconstrained directions. Even in the unconstrained directions, the robot has to manage interaction forces that arise due to friction or deformation. In contact-rich manipulation, we generally do not think of collision avoidance and instead consider the interesting proposition of how to seek and exploit contact. Planning and control of contact-rich manipulation is a challenging problem.

Contact-rich manipulation involves control of the manipulator in contact with the environment. Traditional control schemes are collectively called compliance control [5,6] or *interaction control* [7–9]. Most of these methods assume the availability of a nominal trajectory and deliver fixed or variable compliance behavior along the nominal trajectory. Various studies have shown that such a strategy is an important part of the human manipulation process [10,11]. Ideally, both the motion profile (nominal trajectory) and the compliance profile need to be jointly optimized against the task geometry and also the physical interaction model. Even if the geometry and the interaction models are perfectly known, a joint optimization of the motion and compliance profiles would be nonconvex in general.

In reality, the situation is worse because the geometric model, if at all available, would have inaccuracies and the interaction model is generally unknown. Here, the interaction model refers to a model that describes phenomena such as friction, stiction or deformation along with their associated forces. Thus, except for simple contact cases where a nominal trajectory can be independently generated and a fixed compliance behavior can be assumed, traditional control algorithms are not a general solution for contact-rich manipulation.

A promising alternative to the traditional compliance control approach is robot learning. In robot learning, a control policy is learned that maps state to action, one that can potentially encapsulate both the trajectory and compliance profiles. The policy in robot learning is generally consistent with the concept of skill or behavior. If the policy outputs torques or forces, then there is no longer any explicit trajectory or compliance profile and the policy can be thought of as encompassing the essence of both. If the policy outputs a kinematic variable, such as position or velocity, then further solution for compliance planning and control has to be sought. Therefore, a natural formulation of the policy, in the context of contact-rich manipulation, is one that outputs torques or forces. In robot learning, both Learning from Demonstration (LfD) [12–17] and reinforcement learning (RL) [18–24] have been proposed for contact-rich manipulation. Interestingly, instead of torques or forces, some methods featured policies that produce a combination of trajectory and compliance profiles [23, 24] or a combination of trajectory and force profiles [19, 22]. Since LfD requires expert human demonstrations, that may be inconvenient at times, and also additional sensors (haptic) to register the required forces, RL-based methods may be considered more suitable for contact-rich manipulation.

Although RL has had impressive successes for robotic manipulation in general [25–27] and contact-rich manipulation in particular [18, 20, 23], several aspects of it remain as open problems. Some of the most important problems are:

1. How to guarantee control stability?
2. How to achieve practical sample complexity (data-efficiency)?
3. How to synthesize the reward function?
4. How to achieve domain generalization and domain adaptation?

The first problem can be understood when we realize that the policy learned through RL can be considered as a feedback controller. Since stability is the foremost property that is expected whenever a closed-loop controller is synthesized, it is natural to expect the same for a learned policy. While LfD methods with stability guarantees [15–17] are common, RL-based methods are quite rare. The problem of sample complexity is well-known in RL, but it attains more significance in the context of contact-rich manipulation. This is because random trials in RL that involve repeated contacts and exchange of forces can potentially wear out the hardware. Model-based RL methods [28, 29] are promising in

this regard but those that are tolerant to contact-induced discontinuous dynamics are yet to be demonstrated. The problem of reward synthesis is studied in *inverse reinforcement learning* where the goal is to generate a reward function for a given set of human demonstrations. A notable prior work that includes a real robot demonstration is [30]. The issue of *Domain Generalization* (DG) and *Domain Adaptation* (DA) have been extensively researched in recent years. DA or *few-shot learning* aims to leverage models learned from one task to speed up learning for a new task. DG or *zero-shot learning*, on the other hand, aims to achieve transfer to a new task or domain without any new training. Both DA and DG are usually represented by *meta-learning*; a critical evaluation of the latest methods can be found in [31].

In this thesis, we address the first two of the above mentioned problems in the context of contact-rich manipulation. More specifically, we are interested in answering the following questions:

- **Control stability:**

1. Is it possible to structure a policy such that stability is guaranteed inherently? The motivation being that with an inherently stable policy, existing unconstrained policy optimization methods can be used.
2. How can a robot explore randomly while preserving the stability property?
3. Is it possible to obtain provably stable policies when they are parameterized as deep neural networks?
4. How does imposing stability guarantee affect other aspects of RL? Will it increase or decrease the sample complexity?
5. How to reason about stability when the environment with which the robot is physically interacting is unknown. What assumptions are necessary?

- **Data-efficiency:**

6. If a model-based RL approach is taken to achieve data-efficiency, how to effectively learn dynamics models that feature contact-induced discontinuities?
7. How can prior knowledge about the nature of contact dynamics be used for model learning and motion prediction?
8. Can methods based on dynamics priors lead to data efficiency?
9. How to exploit structure in learned dynamics models for policy search?

1.2 Thesis Contributions

This thesis is a compilation of four papers [32–35]. A detailed summary of these papers is given in Chapter 3. In this section, the included papers are listed along

with a brief description of the scientific contributions. The individual contributions by the author of this thesis are also pointed out. The included papers are:

Paper A:

Stability-Guaranteed Reinforcement Learning for Contact-Rich Manipulation

S. A. Khader, H. Yin, P. Falco, and D. Kragic. In *IEEE Robotics and Automation Letters (RAL)*, 2020

Paper B:

Learning Stable Normalizing-Flow Control for Robotic Manipulation

S. A. Khader¹, H. Yin¹, P. Falco, and D. Kragic, preprint, arXiv:2011.00072. Accepted at *IEEE International Conference on Robotics and Automation (ICRA)*, 2021

Paper C:

Learning Deep Neural Policies with Stability Guarantees

S. A. Khader, H. Yin, P. Falco, and D. Kragic, preprint, arXiv:2103.16432. In submission, 2021

Paper D:

Data-Efficient Model Learning and Prediction for Contact-Rich Manipulation Tasks

S. A. Khader, H. Yin, P. Falco, and D. Kragic. In *IEEE Robotics and Automation Letters (RAL)*, 2020

Learning contact-rich manipulation skills with control stability

Papers A-C are different approaches for attaining control stability in a model-free RL framework. No environment models are learned and no assumptions regarding objects in the environment are made except that they are *passive*. The manipulator dynamics is also not utilized in the policy synthesis except that a gravity compensation is assumed. The manipulator is made passive through control and the overall stability is reasoned based on the theory of *passive interaction* between two passive objects. See Section 2.5 for a detailed explanation. This addresses question 5.

Papers A-C also succeed in parameterizing policies such that the manipulator-environment interaction is inherently stable. The methods rely on Lyapunov’s direct method [36] for stability proof. Since a deterministic framework is used for stability analysis, stable exploration is guaranteed by limiting exploration in the parameter space. This approach is followed in papers A and C. The method in paper B uses action space exploration and therefore does not guarantee stable exploration. The method, nevertheless, does have practical stability properties

¹Equal contribution

even during exploration and ultimately produces a deterministic policy that is fully stable. This addresses questions 1 and 2.

Paper B features a policy that is partially neural network while paper C features a policy parameterization that is almost entirely neural network. The method in paper A uses a policy with an analytical form. Question 3 is, therefore, answered in the affirmative.

Papers A-C indicate that the stability property actually helps reduce sample complexity and thus answers question 4.

Contributions by the author: In papers A and C, the author proposed and formulated the idea, designed, implemented and evaluated the methods, and wrote the vast majority of the paper. All experiments were designed and performed by the author but after receiving the implementation of the baseline methods. In paper B, the author made significant contributions to the method development, wrote the large majority of the paper, and designed and performed the experiments. The author made only a minor contribution to the implementation of the method. In papers A-C, the author conducted real robot experiments after additional implementations.

Data-efficient learning of contact-rich manipulation skills through model-based RL

To achieve the goal of data efficiency, paper D is formulated within the framework of model-based RL. However, the work is limited to only model learning. The focus is on model learning for contact-rich manipulation. To that end, the paper presents a method based on the formalism of *hybrid automata* [37], which is ideal for representing the peculiarities of contact dynamics in robotic manipulation. This answers questions 6 and 7.

Paper D also shows, on the basis of experimental results, that the proposed method can effectively perform motion prediction after learning a forward dynamics model with little data. This answers question 8.

From a skill learning point of view, the most relevant question is whether the hybrid structure of the learned model can be exploited during policy search. However, this (question 9) remains unanswered in this thesis.

Contributions by the author: In paper D, the author proposed and formulated the idea, designed, implemented and evaluated the method, and wrote the vast majority of the paper. All experiments were designed and performed by the author but after receiving the implementation of the baseline methods.

1.3 Thesis Outline

This thesis consists of two parts. The first part is an overview that contains the motivation, background and summary of the papers included in this thesis. The second part consists of the four included papers. In the overview part of the thesis, Chapter 2 provides a discussion of the scientific background of the

work, Chapter 3 provides a summary of the included papers with more details of the contributions, and Chapter 4 concludes with a discussion of limitations and future work.

Chapter 2

Background

In this chapter, we introduce the scientific background for the contributions in this thesis. Section 2.1 introduces the problem of contact-rich manipulation followed by the necessary background in Section 2.2. The concept of skill learning is presented in Section 2.3, along with its formulation as either Learning from Demonstration (LfD) or reinforcement learning (RL). In Section 2.4, the peculiarities of learning contact-rich manipulation skills are considered and a case for RL is made. Finally, in Sections 2.5 and 2.6, the issues of control stability and data-efficiency in RL-based skill learning are examined, respectively.

2.1 Contact-Rich Manipulation

Autonomous robots operating in unstructured environments have to deal with uncertainties. Common sources of uncertainties are errors in modeling, sensing and actuation. Consider, for example, sensing; modern robots are equipped with vision, force, distance and touch sensors and are expected to process these streams of data and build unified internal representations. It is quite natural for uncertainties to creep into the internal models. Autonomous robots in an industrial production environment may also have to deal with tolerances in part sizes.

Uncertainties can have a significant impact on all aspects of manipulation. Consider the example of gear assembly in Fig. 1.2. Uncertainties in the position, size and pose of the gear can have an impact on the grasping process. The same is also true for the collision-free motion planning phase, if uncertainties exist for the locations of the objects in the environment. However, for these two phases: *grasping* [2] and *collision-free motion* planning [3], it is common to bound the uncertainties and plan with a sufficient margin without explicitly taking the uncertainties into account. Unfortunately, such an approach is not possible for the gear insertion phase, where even a tiny amount of uncertainty in the relative location, pose or size of either the gear or shaft would result in a collision. With the classical approach of motion planning and control, where a

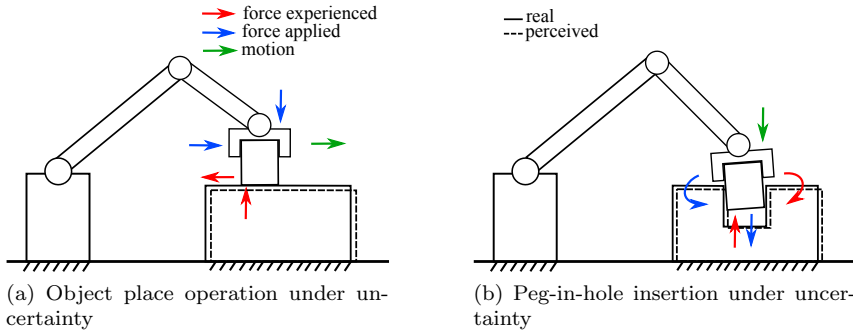


Figure 2.1: Motion and contact forces in contact-rich manipulation

trajectory is planned independently to be tracked by a feedback controller, such collisions can result in large forces and cause damage. In such a scenario, more sophisticated motion generation and control algorithms are required that can not only seamlessly handle unexpected contacts, but also plan based on possible contacts. We refer to this aspect of manipulation as *contact-rich manipulation*.

It may be pointed out that contact-rich manipulation should not be seen as something relevant to insertion or assembly tasks only. Consider a simple pick and place operation where a manipulator has to pick an object and place it on the top of a surface. The place operation is depicted in Fig. 2.1a where the real and perceived locations of the surface are shown. The discrepancy exists due to uncertainty. In this example, it can be concluded that a trajectory planned according to the perceived height of the surface will penetrate the actual surface and thus would result in collision.

More generally, contact-rich manipulation involves *constrained motion* and *compliant motion* [4]. The former refers to the situation where the manipulator motion is constrained by a rigid object. The forces applied by the manipulator are balanced by the reaction from the surface normals. The latter refers to the motion of a manipulator that is in continuous contact. While in contact, the manipulator may slide along a surface and experience frictional phenomena that also give rise to forces. In addition to frictional forces, physical interaction can also subject a manipulator to inertial forces, e.g. pushing a block, and elastic forces, e.g. pushing against an elastic wall.

Notice that in all our considerations, we assume that a stable and rigid grasp is already established, an assumption that will be held throughout this thesis. Moreover, within the context of contact-rich manipulation, as exemplified by the gear assembly task in Fig. 1.2, there shall not be a consideration of collision avoidance; rather, we shall be interested in the possibility that a sophisticated motion generation and control algorithm could seek and exploit contacts in order to eliminate uncertainties during the manipulation process. A simplified task

that represents all the complexities involved in contact-rich manipulation is the peg-in-hole task [6, 20, 38, 39]. See Fig. 2.1b for a two-dimensional illustration. Here, the goal is to insert a rigidly grasped peg into a hole under uncertainty.

2.2 Classical Approaches for Contact-Rich Manipulation

A classical approach to endow an autonomous robot with contact-rich manipulation capability is the *active compliant motion (ACM)* system [4]. It is mainly composed of *fine motion planning*, *compliant motion planning* and *contact state identification*. While fine motion planning [40, 41] refers to the general strategy of planning fine scale motions that take into account contact forces, friction and geometry, compliant motion planning [42] specifically addresses motion planning under continuous contact. Contact state identification deals with monitoring and identifying the exact contact state at any given time. The ACM system is consistent with the 3T architecture that was mentioned in Section 1.1; the first two components of ACM can be seen as the planning layer and the last component as belonging to the execution layer. The behavior layer would execute the compliant motion plan with interaction control methods such as *impedance control* [43] or *hybrid position/force control* [44].

Interaction control methods [7, 8] are thus an important part of contact-rich manipulation. In one of the earliest works, the hybrid position/force control [44] was introduced to simultaneously deal with motion and force aspects. The lack of consideration of manipulator dynamics in the hybrid approach was later addressed in the operational space formalism [45]. Another seminal method was the stiffness control method introduced by Salisbury [46] that allowed to impart a desired stiffness behaviour without the need for a force sensor. The impedance control approach by Hogan [43] can be seen as a generalization of the stiffness control to include also inertia and damping properties to the interaction behavior. An approach that inverts the velocity-to-force causality of impedance control to force-to-velocity causality is *admittance control*. Admittance control is suitable for most industrial manipulators, the majority of which are non-backdrivable. A comparison of both strategies in [47] revealed that while impedance control is better suited for rigid interactions, admittance control is the better choice for non-rigid cases. Finally, an important concept is that of *variable impedance control (VIC)* where the impedance parameters, inertia, damping and stiffness, are varied—instead of being kept constant—according to task requirements. VIC is believed to give rise to rich interaction behaviors [48].

The biggest drawback of the active compliant motion system is its high computational needs in general [4]. The planning algorithm has to process complex geometric and interaction models and even take into account interaction forces. A key feature of such systems is the generation of contact states that is generally intractable except for simple geometries [49]. Even with a perfectly generated contact states and transition model, the system would also require contact state

identification that is generally error prone [4]. Beyond contact state detection, compliant motion planning involves generation of motion trajectory and its associated compliance profiles. For example, the VIC scheme requires a varying impedance profile to be generated according to the interaction properties. Even with accurate interaction models, which is highly unlikely, no general solution exists for the compliant motion planning problem. In general, it is challenging to scale the active compliant motion system to arbitrary manipulation tasks.

2.3 Robot Skill Learning

In this section, a brief introduction to *robot skill learning*, also known as *robot learning*, is given without any particular focus on contact-rich manipulation. Skill learning for contact-rich manipulation is taken up in the next section.

Although robot learning could have a wider interpretation with regard to applying various machine learning algorithms to solve robotics problems, we adopt the particular interpretation of learning control and motion generation [4]. It is desirable to synthesize expressive control behaviors that assimilate, as much as possible, the functionalities of the upper layers in a 3T-like architecture. This allows the upper layers to focus on more general and coarser aspects of manipulation, while a collection of expressive skills in the behavior layer, each of which is reactive in nature with little computational needs, can easily cope with the dynamism and complexity of the task. Recall that the skill, according to the requirements of the behavior abstraction, is a direct mapping from observation to action. Assuming that a sufficiently rich skill is available, its execution is straightforward and computationally cheap. This is the appeal of the skills concept. Of course, synthesizing skills is not trivial, which is exactly the problem that skill learning promises to solve by harnessing the power of machine learning.

Skill learning takes on two main forms in robotic manipulation: *Learning from Demonstration (LfD)* [50, 51] and *Reinforcement Learning (RL)* [52, 53]. In LfD, human demonstrations of a task are used to synthesize a *policy*. In RL, the policy is iteratively improved, based on autonomous trials of the task and a reward function to evaluate its performance, until a good enough policy is obtained. In both cases, the policy is the embodiment of the skill or behavior. Skill learning is generally applied in the context of motion generation after a stable and fixed grasp has been established, although it could potentially include grasping [2] or *dexterous manipulation* [54, 55].

Learning from Demonstration

The human user produces a set of demonstrations, by manually guiding the manipulator to trace out the desired trajectory (*kinesthetic teaching*), which is then fed into a learning algorithm that optimizes the parameters of a policy. The learned policy is expected to be able to generate the desired motion behavior. A

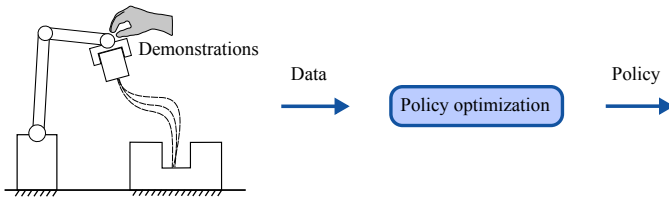


Figure 2.2: Skill learning: Learning from demonstration

survey of LfD methods for robot skill learning can be found in [51]. See Fig. 2.2 for an illustration of the learning process.

Let $s \in \mathcal{S}$ be the state variable and $a \in \mathcal{A}$ be the action variable, where the sets \mathcal{S} and \mathcal{A} are the state and action spaces respectively. Then, the set of N demonstrations, each with an episode length of T , can be represented as $D = \{(s_{0:T}, a_{0:T})^i\}_{i=1}^N$. If $a = \pi_\theta(s, t)$ is the policy parameterized by θ , the LfD learning problem can be summarized as,

$$\min_{\theta} \mathcal{L}(D, \pi_\theta(s, t))$$

where \mathcal{L} is a suitable loss function that measures the error between D and what the policy π would produce. The time dependency of the policy is represented by the variable t . Many LfD methods have either explicit or implicit time dependency although it is not strictly required.

A popular approach to model the policy in LfD is the *dynamical systems (DS)* approach. *Dynamic Movement Primitive (DMP)* [50, 56] is one such representation that can be learned from a small number of demonstrations. It is composed of a global attractor toward the goal position and a motion shaping function that shapes the path the robot takes. A probabilistic version of DMP is the so-called ProMP introduced in [57]. Another instance of DS, that is strictly a function of state, is the approach of modeling a joint distribution of position and velocity as *Gaussian mixture models (GMM)* and then obtaining the policy as a conditional on position through *Gaussian mixture regression (GMR)* [58–60]. An important point to note with regard to the DS approach is that in most cases the learned policy evolves independently in time while generating a kinematic motion profile, usually in the form of velocity. The velocity command is then fed into a low-level proportional-derivative (PD) controller for tracking. In this formulation, the action variable a is in fact the velocity command. The state variable s always includes position but may or may not include velocity.

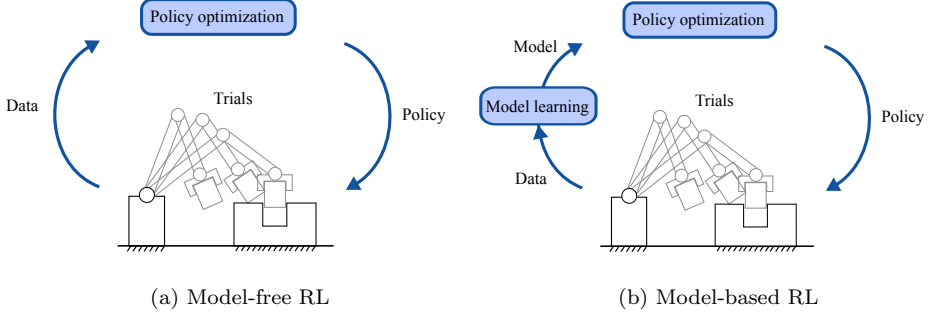


Figure 2.3: Skill learning: Reinforcement learning

Reinforcement Learning

In contrast to LfD, RL promises the possibility of autonomous policy learning¹. Instead of providing the learning algorithm with demonstration data, it is supplied with an appropriate reward function that encourages the desired behavior. The learning process starts with a base policy, often randomly initialized, and is then improved by trial and error in multiple iterations. In each iteration, the robot attempts to perform the task by executing the current policy. The trial phase is marked by some version of random exploration that would potentially discover high reward behavior. Extensive surveys on the topic can be found in [53] and [61]. See Fig. 2.3 for an illustration of the process.

Reinforcement learning is formulated as a *Markov decision process (MDP)* where an agent interacts with an environment to solve a sequential decision making problem. It is formally defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T})$. The sets $\mathcal{S} \subseteq \mathbb{R}^n$ and $\mathcal{A} \subseteq \mathbb{R}^m$ are the state and action spaces, respectively. The agent acts on the environment through the action space and makes observations through the state space. The notation \mathcal{T} represents the transition probability, or dynamics, of the environment and is described by the conditional probability distribution $p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)$, where $s \in \mathcal{S}$, $a \in \mathcal{A}$ and t is the time index. The environment is generally assumed to be unknown. The reward \mathcal{R} is a scalar function, $r(\mathbf{s}_t, \mathbf{a}_t)$, that gives the immediate reward of taking action a_t in state s_t . The solution to the MDP problem is obtained by finding the optimal stochastic policy $\pi_\theta(\mathbf{a}_t | \mathbf{s}_t)$ by maximizing the expected cumulative reward. For an episodic problem with time horizon H , the policy optimization problem can be summarized as,

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{\mathbf{s}_0, \mathbf{a}_0, \dots, \mathbf{s}_H} \left[\sum_{t=0}^H r(\mathbf{s}_t, \mathbf{a}_t) \right],$$

¹Although RL is not necessarily limited to policy search methods, we shall focus on it due to its prominence in robot learning.

where $\{\mathbf{s}_0, \mathbf{a}_0, \dots, \mathbf{s}_H\}$ is a sample trajectory from the distribution induced in the stochastic system and θ^* is the parameter value for the optimal policy.

RL algorithms can be broadly categorized into *model-free* [26, 62–66] and *model-based* [18, 28, 29, 67, 68]. As illustrated in Fig. 2.3, model-based RL algorithms first learn a model of the environment \mathcal{T} , from the data acquired from trials, before optimizing the policy. This is repeated in every iteration. Model-free RL methods, on the other hand, avoid such an intermediate step and directly optimize the policy using the collected data. Model-based methods are known to be more data-efficient [28] but it is at the expense of introducing an additional machine learning problem—*model learning*.

Given a set of N random trials $D = \{(\mathbf{s}_0, \mathbf{a}_0, \dots, \mathbf{s}_H)^i\}_{i=1}^N$, the model learning problem can be formulated as,

$$\min_{\theta} \mathcal{L}(D, p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)),$$

where \mathcal{L} is an appropriate loss function, for example, that maximizes the log likelihood of D . A survey on model learning can be found in [69].

2.4 Learning Contact-Rich Manipulation Skills

In this section, we shall examine the peculiarities of contact-rich manipulation. This is followed by discussions on possible policy requirements and, finally, we conclude by pointing out the preferences one could make for learning contact-rich manipulation skills.

Important features of contact-rich manipulation are:

1. Constrained motion due to unexpected contact and frictional effects
2. Specific forces may be required to accomplish relative motion
3. Contact dynamics is discontinuous in nature

The first feature is due to unexpected contacts with surfaces in the environment that force the manipulator to be constrained in some directions. This cannot be avoided because of the presence of uncertainties in the environment or the robot model. The motion is constrained either due to the blocking action by obstacles or static friction that needs to be overcome. Either way, this gives a discontinuous character to the motion. The second feature emphasizes the fact that unlike free space motion, contact-rich motion often requires the manipulator to deliver task specific forces to the environment to counteract interaction forces. The last feature reminds us that any model learning algorithm would have to deal with the complexities of learning a discontinuous model.

Following from the above features, the most important requirements for a policy could be:

1. **Force/torque as action spaces:** Policies that output only kinematic quantities such as position or velocity are less suitable for contact-rich manipulation. When motion is inhibited due to an unexpected contact or static friction, the low-level controller that tracks the position or velocity command would saturate and enter a fault condition. A force/torque policy would not have any such difficulty.
2. **State-dependent policies:** Policies that generate control action as a function of time instead of state are less suitable for contact-rich manipulation. This is because a manipulator could be blocked by any number of motion constraints in the environment and a time dependent policy would easily get out of sync with reality. A state dependent policy would stay in sync with the dynamics of the environment.
3. **Policy should learn interaction behavior:** Since physical interaction between the manipulator and the environment involves exchange of forces, the policy should be able to deliver the right amount of forces in addition to achieving the right motion profile. The manipulator should comply with environmental conditions when necessary but without generating excessive forces.

Based on the above set of requirements, a set of preferences for skill learning could be listed:

1. **State-dependent policy:** State-dependent policies [15, 17, 70] may be preferred over time-dependent policies such as movement primitives [48, 71].
2. **Force/torque as action space:** Policies that output force/torque [15, 17, 18] can be preferred over the ones that output position or velocities, such as [58–60].
3. **VIC as action space:** Another alternative is to adopt the structure of VIC (see Section 2.2) as the action space. The policy would output the desired position as well as the varying stiffness and damping gains. Using these quantities and the measured position and velocity, a VIC controller can deliver the force/torque to the manipulator. Examples of VIC-based policies are [16, 17, 23, 24, 48, 70]
4. **RL-based skill learning:** RL approaches such as [18, 20, 23, 48, 70] may be preferred over LfD since contact-rich manipulation requires the manipulator to deliver task specific forces to the environment. Demonstrating desired forces is arguably more complex than demonstrating motion trajectory and would incur additional costs in the form of haptic sensors or data gloves. An example of an LfD work that did succeed in such a demonstration is [14].
5. **Deep RL:** A deep RL approach such as [18, 20, 21, 23, 68, 72, 73] would be able to harness the expressivity of deep neural network policies. This is likely to aid

complex behavior synthesis that includes motion generation and interaction control.

To conclude, an RL-based skill learning approach where a deep neural network policy is designed to be independent of time and features an action space that is either force/torque or variable impedance parameters (VIC) is ideal for contact-rich manipulation.

2.5 Control Stability in Skill Learning

Stability is the first property to guarantee whenever a closed-loop feedback controller is synthesized. Control theory is rich in tools with which to analyze and design stable feedback controllers, for both linear systems and complex nonlinear systems. Most existing methods synthesize controllers with analytic structure based on an available dynamics model of the controlled system. Of particular interest to manipulator control is the Lyapunov stability [36] analysis, also the main method for nonlinear systems in general. Stability analysis for various manipulator control problems are well-established and an introduction to the topic can be found in [7]. The concept of policy, that has been used in this thesis to embody the notion of skills, can be seen as a closed loop feedback controller in the regulator sense. Therefore, it is only natural to expect learned policies to conform to the standard notion of stability. A Lyapunov stable RL would guarantee convergence of motion towards a desired goal position irrespective of exploration and the extent of policy training. This would naturally provide predictability, and some amount of safety, to the entire process.

Lyapunov Stability

In Lyapunov stability analysis, an *equilibrium point* of a nonlinear dynamical system is *stable* if the state trajectories that start close enough remain bounded around it. If, in addition, the state trajectories eventually converge to the equilibrium point, then it is said to be *asymptotically stable*. If the system only has a single equilibrium point, then one can refer to global stability of the system instead of any particular equilibrium point.

The precise mathematical definition of the Lyapunov stability method is as follows. Consider an *autonomous nonlinear system* [36], represented by the *differential equation* $\dot{s} = f(s)$, where $s \in \mathbb{R}^d$ is the state variable. Let $s = 0$ be an *equilibrium point* and $D \subseteq \mathbb{R}^d$ be a region that contains the origin. Let $V(s)$ be a *continuously differentiable* scalar function. Then, $s = 0$ is *stable* if,

1. V is *positive definite* in D , or $V(0) = 0$ and $V(s) > 0 \forall s \in D \setminus 0$
2. \dot{V} is *negative semidefinite* in D , or $\dot{V}(s) \leq 0 \forall s \in D \setminus 0$

If, in addition,

3. \dot{V} is *negative definite* in D , or $\dot{V}(s) < 0 \forall s \in D \setminus \mathbf{0}$,

then $s = 0$ is *asymptotically stable*.

Furthermore, if,

4. $D = \mathbb{R}^d$ and

5. V is *radially unbounded*, or $\|s\| \rightarrow \infty \implies V(s) \rightarrow \infty$,

then $s = 0$ is *globally asymptotically stable*.

The nonlinear system $\dot{s} = f(s)$ is an abstract *autonomous system*. In the case of a controlled dynamical system, the corresponding autonomous system is formed as $\dot{s} = f(s, \pi(s))$ where $\dot{s} = f(s, a)$ represents the system dynamics and $a = \pi(s)$ represents the feedback controller. Note that the formalism above is for a deterministic system, unlike the stochastic formulation of RL in Section 2.3. The result generalizes to any equilibrium point in \mathbb{R}^d through a simple translation transformation of the state variable.

Learning Manipulation Skills with Stability-Guarantee

In the field of skill learning, the DMP policy parameterization, in its original form, has an asymptotic convergence property towards the goal. LfD methods [50, 56, 57] and RL methods [74–76] that are based on DMP inherit this property. However, DMPs are often formulated as time-dependent trajectory generators that depend on low-level controllers to track the generated trajectory. Although a case can be made for the overall stability of the system—if the gains of the low-level controller are kept fixed—such a solution is ill-suited for contact-rich tasks. To remedy this, stable dynamical systems were formulated in [15–17] to be time-independent policies that directly output torque or forces. While these methods were essentially LfD, Rey et al. [70] formulated an RL solution along these lines but without establishing complete stability. See paper A for a discussion. Therefore, an RL-based solution for learning contact-rich manipulation skills that guarantee stability, especially with neural network policies, is not only an open problem but hardly any work exists to date.

Challenges in Stability-Guaranteed RL

The difficulty in realizing stability-guaranteed RL can be seen from Fig. 2.4. In this figure, a general possibility is sketched for the purpose of discussion. First of all, the fact that the dynamics model is assumed to be known in the Lyapunov analysis is respected by formulating a model-based RL approach. In model-based RL, the dynamics model is not known to begin with and any learned model is updated in every iteration. This implies a possibility that the Lyapunov function itself is learned. Would the Lyapunov function synthesis be based on the learned

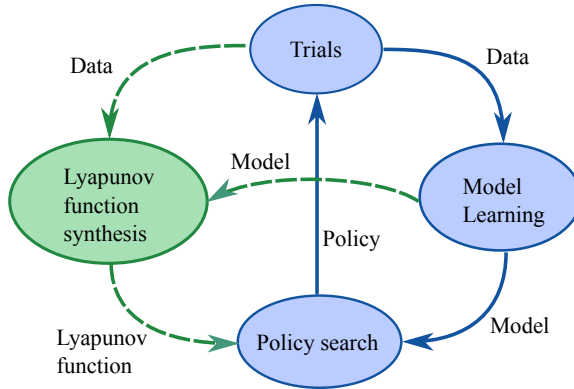


Figure 2.4: A possible framework for stability-guaranteed model-based RL

model or directly from data? How to optimize the policy using both the model and Lyapunov function? How to guarantee stability during random trials? Would it be possible to have a model-free RL algorithm in practice? If so, would it be possible to parameterize a particular policy and then deploy a state-of-the-art model-free RL algorithm? Is it possible to apply Lyapunov stability analysis on deep neural network policies? These are some of the difficult questions that arise in this context.

The closest method to the model-based RL approach is [77] except that the Lyapunov function is not learned. The question of learning the Lyapunov function is studied in [78] but outside the framework of RL. Furthermore, the method in [77] would struggle to cope with contact-rich manipulation due to its reliance on smooth Gaussian process (GP) [79] model of the dynamics. Another GP based method [80] to learn dynamics is further limited to learning only the unactuated part of the dynamics with the assumption that the remaining part is known. The method in [81] requires a stabilizing prior controller, the uncertainty of which together with that of the learned model determines the region of guaranteed stability. Ideally, it would be good to avoid learning complex dynamics of robot-environment interaction and also be free of any requirements of prior stabilizing controllers. Furthermore, a solution that achieves stability based on only policy parameterization would benefit from state-of-the-art model-free policy search, greatly simplifying practical RL-based skill learning.

Stability through Passive Interaction

An important property of the manipulator-environment interaction process is the passive interaction property. If the manipulator is made *stable* and *passive* with respect to the energy port $(F_{\text{ext}}, \dot{\mathbf{x}})$, where F_{ext} represents the force variable and $\dot{\mathbf{x}}$ represents the velocity variable, then any interaction with a passive environ-

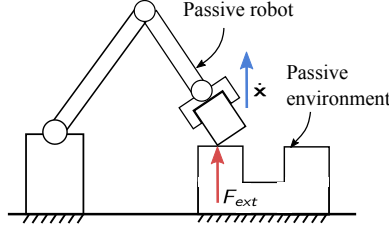


Figure 2.5: Passive interaction between the robot and the environment. F_{ext} represents the external force acting on the manipulator and $\dot{\mathbf{x}}$ represents the end-effector velocity.

ment through this port will result in a stable coupled system [9, 82] (Fig. 2.5). The significance of this property is that, if the environment is passive, then it is enough to establish stability (and passivity) of only the manipulator in isolation. This provides a tremendous opportunity since it removes the necessity of learning the interaction dynamics; in fact, no model learning is required because the manipulator dynamics is well-known and need not be learned. This opens up the possibility of a model-free RL approach.

Passivity of a system is reasoned as follows. Consider a *dynamical system* that is influenced by an unknown external input u and has an output y . The system

$$\dot{s} = f(s, u) \quad y = h(s, u)$$

is *passive* if there exists a *continuously differentiable* lower bounded function $V(s)$ with the property $\dot{V}(s) \leq u^T y \forall u, y$. In the case of a manipulator, f is the dynamical system composed of the control law and the manipulator dynamics, $u = F_{\text{ext}}$ is the force experienced during interaction, $y = \dot{\mathbf{x}}$ is the velocity and h is any appropriate function. The variable s is the state as defined earlier. Passivity means that a system can only dissipate or store energy and not generate it. This is inherently true for unactuated objects in the environment due to the law of conservation of energy.

In this formulation, the only remaining question is how to model the parameterized policy and the Lyapunov function for a manipulator moving in free space such that it is stable and passive. It may be pointed out that the requirement of a passive environment is not limiting since all it means is that the environment is not actuated. Thus most objects in the environment are passive and the obvious exceptions are other robots and humans.

2.6 Data-Efficiency in Skill Learning

Reinforcement learning has an advantage that it is possible to learn complex control policies without having to model the environment. When the environment dynamics is difficult to model, such as the case in contact-rich manipulation, RL

algorithms shine in their utility. However, one of the most worrisome concerns regarding RL is its possible requirement of a large number of trials. The expected number of trials, or *samples*, of an RL algorithm is called its *sample complexity*. For instance, the groundbreaking DQN method [83] for playing Atari games used ten million frames for training. An order of magnitude reduction in training steps was achieved in DDPG [64] and NAF [26] methods that focused on continuous control tasks. However, to achieve practical training times, in real-world application of RL, and also to minimize wear and tear of physical systems, the total number of data samples for training has to be brought down to at least thousands.

Model-based RL is considered as a promising approach for data-efficient policy learning [84]. As shown in Fig. 2.3b, model-based RL methods learn a model of the dynamics as an intermediate step and use it to optimize the policy. It was shown in a number of works [28, 29, 85, 86] that a probabilistic model learning and uncertainty propagation approach can significantly reduce sample complexity down to hundreds of trials. Therefore, model-based RL offers a practical solution for real-world robotic manipulation tasks, especially those that can benefit from reduced physical interaction.

Model Learning

As mentioned in Section 2.3, the model learning step in model-based RL is an independent machine learning problem. Here, the model of interest is the forward dynamics model of the environment. The model can be used to predict the trajectory of the system, which can then be used to evaluate the underlying policy. A policy that produces high cumulative reward, based on the system trajectory, is preferred to the one with a lower reward. In the seminal work PILCO [28], Deisenroth et al. showed that a probabilistic model that represents both *epistemic*² and *aleatoric*³ uncertainties along with a *long-term prediction* model that propagates these uncertainties can drastically reduce the amount of training data. In [28] (PILCO) and [85] (GP-MPC), Gaussian process (GP), that inherently includes both types of uncertainties, is used to learn models and *moment matching* is used to perform long-term predictions. In [29] (PETS), an *ensemble of bootstraps* approach combined with a particle based method is used to achieve the same results using neural networks. Deterministic model learning and long-term prediction using deep neural networks was done in [67] and [68].

Discontinuous Dynamics in Contact-rich Manipulation

In the context of contact-rich manipulation, the environment in the RL sense is in fact the coupled system of the physically interacting manipulator and the object. As noted in Section 2.4, learning dynamics models in this case can be challenging

²Epistemic uncertainty is the uncertainty due to the lack of training data.

³Aleatoric uncertainty is the uncertainty due to the inherent randomness in a system.

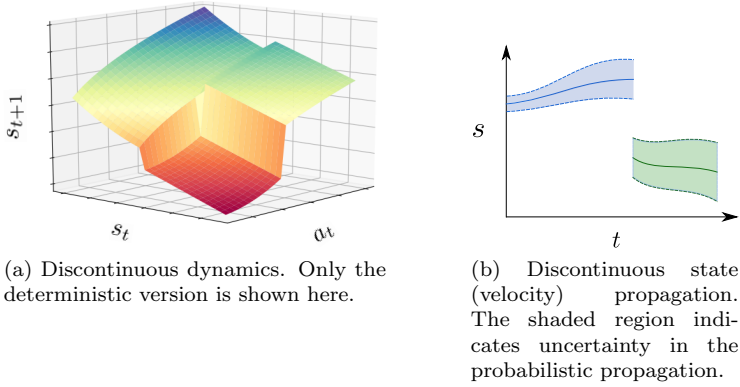


Figure 2.6: Illustrations of discontinuous dynamics and state propagation

due to its discontinuous nature. Recall that discontinuity arises due to collision with constraints and rapid making and breaking of friction between the contact surfaces. In addition to discontinuities in the forward dynamics function, contact can also cause discontinuous transitions in velocity. Multi-step long-term prediction using a learned dynamics model should be able to faithfully reproduce such discontinuities in state propagation. Finally, for reasons mentioned earlier, both the model and the long-term prediction should be probabilistic in nature. Figure 2.6 shows an illustration.

Most machine learning models, be it GP or neural network, have an underlying assumption of smoothness between two data samples. Normal Gaussian process regression (GPR) will have difficulty in distinguishing between a discontinuity and noise. Regular neural network regression would require complex models and large amounts of data to approximate a complex function such as discontinuity, thereby defeating the purpose of model-based RL. Despite these difficulties, several methods employed common modeling techniques without any special considerations for discontinuities. For example, GMM [18], neural network regression [67, 68, 87] and GPR [28, 85] have been used for learning predictive models. Non probabilistic state propagation was done in [67, 68].

Two notable methods that increase the expressivity of the learned model while maintaining probabilistic representation of the model and probabilistic long-term prediction are [88] and [29]. The increased expressivity is expected to help model discontinuities in the learned model but no attention was given to discontinuous long-term prediction. Manifold GP [88] introduced neural network feature mapping before applying the squared exponential kernel function in a GP. The parameters of the mapping and the regular GP hyperparameters were jointly optimized. The work only featured one-step prediction and therefore did not include long-term prediction. PETS [29] introduced an ensemble of bootstrap approach to learn neural network models with uncertainty representation, something the

normal GP and also the manifold GP had built in. It also presented particle based long-term probabilistic prediction. However, none of these works explicitly validated long-term prediction using a learned model for contact-rich tasks; instead, they only validated the overall policy learning.

Learning Hybrid Models

An appropriate theoretical construct that models discontinuous dynamics well is *hybrid automata* [37], which is a special member of the general family of hybrid systems. Generally, a hybrid system is characterized by discrete *modes*, each of which represents a smooth model. In the case of the dynamics model depicted in Fig. 2.6a, each of the smooth regions would be called a mode, and an instantaneous transition between two adjacent modes would represent a discontinuity. Learning methods for hybrid systems learn each of the modes and also a selector function [89] [90] based on the current inputs. This is related to the *mixture of experts* approach in learning expressive mixture model [91], except that there is no soft mixing but only a hard switching. More sophisticated approaches such as [92] and [93] also consider the current mode, in addition to the current inputs, for the selector function. However, none of the existing methods in learning hybrid models include a solution for discontinuous state propagation.

Summary

In this chapter, we presented the scientific background of this thesis. We first clarified the meaning and scope of contact-rich manipulation and then justified the need for skill learning by RL. Two topics of concern with respect to skill learning in contact-rich manipulation are identified: control stability and data-efficiency. This is followed by more in-depth background and literature review on these topics. Papers A-C deal with control stability in RL and paper D addresses data-efficiency in model learning, an integral part of model-based RL.

Paper A [32] delivers a stability-guaranteed RL method with a VIC-based policy; the policy, which is of an analytic form, is adopted from the prior work [17]. Paper B [33] presents a solution for parameterizing a partially neural policy with stability property but without guaranteeing stability during random exploration. Paper C [34] introduces a deep neural policy with inherent stability and also demonstrates a policy search with complete stability guarantee. All the three works exploit the passive interaction property in order to enable a model-free RL approach.

In paper D [35], we propose a hybrid system learning method based on the formalism of hybrid automata. This formalism has a unique concept called the *reset map* that explicitly deals with the issue of discontinuous state propagation during the long-term prediction. The main contribution is a solution that not only learns discontinuous dynamics models but also performs discontinuous long-

term prediction with the learned model. Additionally, the method is completely probabilistic and hence meets all the requirements mentioned so far.

Chapter 3

Summary of Included Papers

In this chapter, the papers included in this thesis are summarized and their scientific contributions highlighted. The contributions of the author of this thesis are also listed.

Paper A - Stability-Guaranteed Reinforcement Learning for Contact-Rich Manipulation

Summary

In this paper, we address the lack of stability guaranteed RL algorithms for learning contact-rich manipulation skills. Recognizing the importance of VIC in interaction control theory, a number of RL methods were proposed that adopted a VIC structured policy parameterization. However, these methods either did not address stability at all [23, 48] or did so only partially [70]. To convey the scope of the stability guarantee of our method, we introduced the term *all-the-time-stability* that explicitly meant that every possible trial during the RL process will be stability guaranteed. The aim was to develop an RL method with *all-the-time-stability* property.

The proposed solution is crafted based on the requirements that were outlined in Sections 2.4 and 2.5, most of which were already satisfied by the adopted motion modeling framework i-MOGIC [17]. Specifically, the i-MOGIC policy is parameterized in a state-dependent form and features a VIC structure. It also has stability properties subject to certain constraints on its parameters. It utilized the stability property of passive interaction between the manipulator and its environment. With the i-MOGIC policy already satisfying all of our requirements with regard to policy parameterization and inherent stability property, our focus was diverted to the model-free RL aspect. To this end, we introduced a novel gradient-free policy search algorithm that is inspired by Cross-Entropy Method [94] to optimize the parameters of the policy. Our solution for policy

search was such that stability was inherently guaranteed despite an unconstrained search.

The method was validated on a series of simulated two-dimensional block insertion tasks and also a 7-DOF manipulator arm performing a peg-in-hole task. The results confirmed the feasibility and usefulness of stability guaranteed RL. Particularly, we showed that stability guarantee did not come at the expense of sample efficiency. As a part of our study, we reported the first successful stability guaranteed RL that was demonstrated on the standard benchmark problem of peg-in-hole. A limitation of our work was that the policy (i-MOGIC) is of an analytic form and is arguably less expressive than a deep neural network policy.

Contributions

The scientific contributions in this work are:

- We present a solution for stability-guaranteed RL of contact-rich manipulation skills.
- We introduce a novel *evolution strategies* [95]-based policy optimization algorithm closely resembling the Cross-Entropy Method. In particular, our method can handle positive definite matrices, in addition to real-valued vectors, as part of the decision variables.
- We demonstrate, to the best of our knowledge, the first stability guaranteed RL of the peg-in-hole task.

Contributions by the author

- Proposed and formulated the problem.
- Designed and implemented the method.
- Designed and performed all experiments after receiving baseline method implementation.
- Wrote the large majority of the paper.

Paper B - Learning Stable Normalizing-Flow Control for Robotic Manipulation

Summary

Stability guarantee is a desirable property to have in deep RL algorithms for continuous control problems. To benefit from the rapid progress in deep RL research, one would like to impart the stability property to the policy alone and stay within the framework of popular policy search algorithms. One of the reasons why this is difficult is the uninterpretable nature of deep neural network policies. In the context of contact-rich manipulation, an additional challenge is to guarantee stability during physical interaction of a manipulator with an unknown environment. Since no such solution exists to date, we contribute towards closing this gap.

We present the *normalizing-flow* control structure, a deterministic policy that is partly parameterized as a deep neural network and partly with an interpretable *spring-damper* system. It is well-known that a fixed spring-damper system acting as a regulator on a manipulator has stability properties. It is also known that a spring-damper policy would be very limited. Instead of directly controlling the manipulator using such a policy, a 'normal' spring-damper system is set up in a latent coordinate system which is then mapped, bijectively, to the actual coordinate system. This bijective (invertible) transformation function is parameterized as a deep neural network. The control force generated by the 'normal' spring-damper system is transformed into the actual coordinate system by employing the principle of virtual work. By learning only the nonlinear invertible transformation through RL, it is proven that the original stability property, in the sense of Lyapunov, is retained for any parameter value of the mapping. Furthermore, stable interaction with the passive environment is also proved. Our method is inspired from the concept of *normalizing-flow* [96] that is used for density estimation in machine learning.

The method was validated using a simulated block insertion task and also a real-world gear assembly task by a 7-DOF manipulator. We used a state-of-the-art deep RL policy search method despite the fact that the formal stability guarantee would be lost due to the introduction of action space exploration. Nevertheless, the results clearly showed stable behavior even for moderate amounts of exploration noise. Our results also showed that it was possible to achieve exploration efficiency by virtue of the underlying stability property where all trajectories are directed towards the goal. Therefore, not only did our method help bring stable behavior, but also reduced sample complexity. The proposed method showcased how to impart stability behavior by virtue of only policy parameterization while allowing state-of-the-art policy search methods.

Contributions

The scientific contributions in this work are:

- We present a deterministic policy parameterization for RL, called the *normalizing-flow* control, that has inherent stability properties even when interacting with an unknown (passive) environment.
- Our method provides an instance of interpretable neural network policy with provable stability property.
- We show empirically that the stability behavior leads to sample efficiency.
- We show an instance of how to impart practical stability in state-of-the-art RL methods.

Contributions by the author

- Made significant contributions to the problem formulation and method development.
- Made minor contributions to the algorithm implementation.
- Designed and performed all experiments.
- Wrote the large majority of the paper.

Paper C - Learning Deep Neural Policies with Stability Guarantees

Summary

We follow the same motivation as paper B in seeking stability guarantee in deep RL algorithms for robotic manipulation tasks. The difficulty of achieving stability in this scenario can be attributed to three factors: uninterpretability of neural network policies, random exploration and unknown dynamics. By assuming the knowledge of manipulator dynamics and leveraging the stability of passively interacting bodies, the problem of unknown dynamics is largely eliminated. In this paper, the first two aspects are addressed with the goal of having a policy parameterization that is almost entirely a neural network. A policy that is almost entirely a deep neural network would be capable of learning expressive policies.

Our solution is a policy parameterization that is entirely a deep neural network except for an additive component that is linear in the position variable. The additive component is also learned in the process. The policy is derived from physics-based prior of *Lagrangian mechanics*, which has been previously applied for controlling mechanical systems in the form of *energy shaping (ES)* control [97]. The ES control form leads naturally to a Lyapunov function that allows straightforward stability proof. Although the ES control form is well established, how to parameterize a learnable neural network policy is not obvious. The energy shaping policy has a general structure that consists of the gradient of a convex potential function—a function of the position variable—and a nonlinear function of the velocity variable that satisfies a certain property. We use the recently proposed *Input Convex Neural Network (ICNN)* [98], operated on by automatic differentiation to get the gradient, and an appropriate nonlinear function of the velocity to parametrize a deep neural network policy to satisfy the properties of the ES control form. Since stability is now guaranteed, even for randomly initialized networks, we proceed with the Cross Entropy Method to implement a parameter space search for the optimal policy without losing stability.

We validate the method using simulated block insertion tasks and a peg-in-hole task performed by a 7-DOF manipulator. Our results confirm the previous observation made in papers A and B that stability in fact improves sample efficiency due to the convergent behavior toward the goal position. We also show that, due to global stability, our method is robust to random perturbation to initial position after learning. The proposed policy form (ES policy) is compared to the policy in paper B (NF policy) and is found that being almost entirely a neural network does have an advantage. It turned out that the spring-damper system in the NF policy has to be properly initialized, while the ES policy, being fully learnable, has no such hyperparameter to set. Apart from this crucial difference, unlike the work in paper B, stability was guaranteed even during random exploration. This work is thus comparable to paper A but with the important difference that the policy in paper A was of analytic form and thus, arguably, less

expressive.

Contributions

The scientific contributions in this work are:

- We present a complete stability guaranteed RL of learning manipulation skills with deep neural policies.
- Our method provides another instance of interpretable neural network policy with provable stability property.
- We show empirically that the stability behavior leads to sample efficiency.

Contributions by the author

- Proposed and formulated the problem.
- Designed and implemented the method.
- Designed and performed all experiments after receiving baseline method implementation.
- Wrote the vast majority of the paper.

Paper D - Data-Efficient Model Learning and Prediction for Contact-Rich Manipulation Tasks

Summary

In this paper, we focus on the problem of learning a forward dynamics model and using it to perform forward recursive prediction, or *long-term prediction*, in the context of contact-rich manipulation. As explained in Section 2.6, the central challenge is accommodating discontinuities during model learning and also long-term prediction. Model learning and long-term prediction are both necessary steps in model-based RL, which is seen as one of the main approaches to gain data-efficiency. The need for data efficiency is extremely important for contact-rich manipulation since fewer trials means lesser wear and tear for the physical system. While several methods exist that can learn discontinuous models, none exist that can also perform discontinuous long-term prediction. Recall that discontinuity in the predicted state evolution is due to the fact that velocities can change abruptly upon contact. To complicate matters, both model learning and long-term prediction are required to be probabilistic in nature if data efficiency is to be achieved.

We present a solution to learn probabilistic forward dynamics models with discontinuities and also to perform probabilistic long-term state trajectory prediction with discontinuities. To represent a discontinuous model, we adopt the hybrid systems formalism of *hybrid automata* [37], that includes concepts such as *mode dynamics*, *guard function* and *reset maps*. Modes represent the actual dynamics in the smooth regions of the hybrid system. A global guard function learns to predict the next mode based on the current state-action pair. Finally, for each mode transition present in the data, a reset map learns to predict the post transition state based on the pre-transition state-action pair. The modes and reset maps are learned as probabilistic models using GPR while the global guard function is learned as a deterministic model using a *support vector machine*. A long-term prediction algorithm propagates uncertainties through all the learned models using a particle based approach. An interesting feature of our method is that it supports propagation of multimodal state distribution. This would correspond to distinct possibilities that the motion could evolve into in a contact-rich environment.

We evaluate the method using a simple contact motion experiment. RL trials are simulated by perturbing the parameters of a trajectory-centric policy and the dataset collected is used to learn a model. Long-term prediction is performed using a holdout policy and is validated against the corresponding holdout data. Comparison to state-of-the-art baseline methods, *manifold GP* [88] and *PETS* [29], showed that our method is able to perform significantly better under scarce data. A drawback of our method is its low scalability. This is inherited from GPR but parallelization is expected to alleviate this problem.

Contributions

The scientific contributions in this work are:

- We present a method for learning a probabilistic hybrid dynamics model. The main novelty is that the learned model conforms to the structure of hybrid automata. Of particular interest is the concept of reset maps, which allows discontinuous long-term prediction, something that has not been achieved in a learning-based approach prior to our work.
- We showed that data efficiency is increased if the hybrid automata structure is adopted, in addition to probabilistic models, for learning contact-rich motion dynamics.

Contributions by the author

- Proposed and formulated the problem.
- Designed and implemented the method.
- Designed and performed all experiments after receiving baseline method implementations.
- Wrote the vast majority of the paper.

Chapter 4

Discussions and Conclusion

This thesis addresses the problem of skill learning for contact-rich manipulation tasks. The data-driven approach of skill learning is identified as an important component of autonomous robots. The skill learning problem is formulated and studied as a reinforcement learning problem with a focus on control stability and data-efficiency. Both of these aspects are extremely important for real-world application of reinforcement learning on robotic manipulation tasks. In this chapter, we summarize the important conclusions from the thesis contributions and expand on remaining challenges and potential future work.

Despite the recognition that stability is a necessary requirement for control synthesis, almost no solutions exist today for RL-based skill learning when time independence and force/torque based action space are required. Our papers A-C [32–34] contribute towards closing this gap. We showed how to reason about stability in RL despite two difficult obstacles: uninterpretable nature of deep neural network policies and unknown interaction dynamics. Interestingly, the stability property helped to improve sample efficiency in RL.

Data-efficiency is a widely researched topic in RL. In paper D, we adopted the model-based RL framework to address data-efficiency and limited our work to only learning forward dynamics models in the context of contact-rich manipulation. We investigated the possible benefits of adopting a hybrid systems formulation of contact dynamics and found that it indeed helps in achieving accurate model learning under scarce training data. Our work highlighted the need for modeling discontinuities in state prediction in addition to discontinuities in the dynamics model. A learning based method that did both did not exist prior to our work.

4.1 Control Stability

In this section, we examine some remaining challenges and potential future work with regard to control stability.

Model capacity in stable RL

Papers A-C present different strategies for parameterizing policies for RL that suit well for contact-rich manipulation skills. In paper A, a motion modeling framework with an analytic form [17] is used. In paper B, the policy consists of a fixed spring-damper structure and a special deep neural network that performs as a bijective mapping. In paper C, the policy consists of a deep neural network function that is convex in its input (ICNN) [98] and a fully connected network structured as a positive definite matrix. The special parameterizations of the three policies are necessary to prove Lyapunov stability but their effects on the expressivity of the policies are not obvious.

Although our work includes a limited comparative study between models in papers B and C, a more comprehensive study that not only compares the three models, but also with a regular deep neural policy is recommended. Such a study should examine theoretical limitations on motion profiles that can be generated. At the very least, empirical studies may be undertaken to establish the modeling capacity of the policies with the help of a wider range of manipulation tasks.

Exploration in stable RL

One of the limitations in our work is that exploration is limited to only the parameter space if stability is to be preserved. In papers A and C, we use the Cross-Entropy Method (CEM) [94]-based approaches for parameter space exploration, which is a gradient-free black-box optimization method. In paper B, we use the more common action space exploration but lose the stability guarantee in the process. Note that the final deterministic controller would still be stability guaranteed. Exploration in action space can permit the application of more efficient policy gradient RL methods while exploration in parameter space with deterministic policies only allows less efficient gradient-free methods such as CEM, CMA-ES [99] or NES [100].

A straightforward remedy is to use scalable versions of CEM-like algorithms, for instance the approach in [101]. A more elegant alternative is to extend the deterministic Lyapunov analysis to a stochastic version [102] where action space exploration could be permissible. This is a promising future direction that could harness the efficiency and scalability of state-of-the-art policy gradient algorithms such as [65, 66, 103].

Generalizing stability

In this thesis, stability is defined in the sense of Lyapunov’s direct method [36] and is used in reasoning about stability of a passive interaction between a manipulator and its environment. Thus, stability only holds for passive objects in the environment. Although this is not a serious limitation, as argued in Section 2.5, it may be noted that with active objects in the environment, such as other

robots or humans, stability is not guaranteed. It may be necessary to adopt other strategies to cope with such situations [104].

Although our methods do not promise safety, many recent works in safe RL propose safety through the concept of Lyapunov stability [78, 105, 106]. Stability promotes safety only to the extent that certain guarantees exist that ensure state evolution (trajectory) toward a known goal position. If a human or any delicate object gets in the way of a stable robot, there are no guarantees on bounds on the force that the robot may apply. A robot with sufficient energy could do serious damage in that situation. Therefore, an interesting future direction could be to adopt suitable notions of safety that may even subsume stability.

In this thesis, we have formulated the underlying Markov decision problem such that the state variable consists of robot position and velocity. For a manipulator arm assumed to be a rigid body, such a formulation is appropriate. However, if the state is not fully observed, for example if the observation is in the form of a visual input, stability analysis could become more complex. See [107] for an example. Achieving the same result in an RL-based skill learning framework is highly desirable.

4.2 Data-Efficiency

In this section, we examine some remaining challenges and potential future work with regard to data-efficiency.

Policy search with learned hybrid models

In order to achieve data-efficiency through model-based RL, solutions are needed for both model learning and policy search. We have addressed only the first part—model learning. A straightforward approach would be to use the learned hybrid model as a black-box simulator and use existing model-free policy search methods. This could also include a (gradient-free) model predictive control (MPC) [108] approach. More interesting alternative would be to exploit the structure of the learned models. For example, the presence of a number of discrete modes in the hybrid model could be synergized with sub-policies (or *options*) within a hierarchical RL approach such as the *option-critic architecture* [109]. A partially successful attempt is reported in [110]. Future improvements could also investigate how to exploit the multimodal uncertainty propagation capability of our method.

Learning hybrid systems

Our hybrid dynamics model learning method has some limitations. In most learning algorithms that learn switched system models [92, 93], the clustering part and mode learning part are not independent but are intertwined. In our work (paper D), we assume linear separability in a feature space and perform the

clustering step independently followed by model learning. The former approach is considered more optimal but not directly applicable due to the uniqueness of our formalism (hybrid automata). This could be addressed in a future work. A closely related issue is that we employ iid clustering when in reality the dynamics data are temporally correlated. This could be improved by applying trajectory segmentation methods to consider the temporal correlations.

Scaling model learning

Our model learning method, being based on GPR, would face difficulty in scaling to a large training dataset. Despite this, GPs are attractive due to its data efficiency and also their inherent capability to represent both epistemic and aleatoric uncertainties (see section 2.6). A straightforward remedy is through massive parallelization for which our method is especially suitable. Along this line is also the possibility to scale up a potential gradient-free policy search method; see [111] for an example.

4.3 Skill Learning

The two problems studied in this thesis, control stability and data-efficiency, are treated independently. Furthermore, the approaches are based on fundamentally different RL formulation; stability-guaranteed RL is formulated as model-free RL and data-efficient RL is formulated as model-based RL. We provide a recommendation for when to use each of them. We also discuss the prospects of unifying both problems in one RL formulation. Finally, we touch on the important topic of generalization in skill learning.

Stability in model-based RL

The most obvious approach to unify stability guarantee and data-efficiency is to aim for stability-guaranteed model-based RL. A prominent example of such a solution is [77] where the model is learned through GPR and stable exploration is performed according to a given Lyapunov function. A general solution may also involve the additional step of improving the Lyapunov function [78]. In general, a model-based stability-guaranteed RL would be computationally expensive; for example, the method in [77] requires discretization of the state space and subsequent point wise evaluation. With hybrid models involved in the process, the complexity of such an approach is only going to increase significantly. A survey on stability analysis for hybrid systems can be found in [112].

Model-based or model-free?

Our work on model learning for contact-rich manipulation revealed that learning hybrid systems, especially one that is modeled as a hybrid automata, is signifi-

cantly more complex than a naive regression process with neural networks or GPs. We also discussed the potential difficulties with a stability-guaranteed model-based RL. Given this information, the pertinent question is whether a model-based RL approach is advisable for learning contact-rich manipulation skills at all.

Based on the results in this thesis, we recommend a model-free approach as the first choice for stability-guaranteed RL. In this regard, our results are encouraging since it indicates that the stability property itself has a positive effect on improving data-efficiency. However, the model-based RL approach is still relevant for a number of reasons. First, it has the potential to give the best sample efficiency, which is important in real-world application of skill learning. Second, a model-based RL in the form of MPC opens the door for elegant unification of planning and learning approaches. In an MPC-based RL approach [29, 67, 85], the model is first learned and, instead of learning a policy, an online optimization process determines the control action. The online optimization process can potentially incorporate stability guarantee, safety and resource constraints, or any other planning process that cannot be solved based on a learning paradigm. Therefore, a model-based MPC framework would offer a remarkable capability for an autonomous robot, something that is difficult to achieve in a purely policy learning framework.

Generalization in skill learning

Generalization to novel situations is one of the main limitations of the skill learning approaches. In this regard, the traditional active compliant motion systems discussed in Section 2.2, if designed appropriately, would have an advantage over skill learning systems. The planning modules of such systems would simply take in the geometric and interaction models of the new task, or domain, and instantly be ready for operation. In contrast, most learning based systems would require additional training sessions to adapt to the new task. Still, skill learning may be preferable due to its practical viability and its inherent ability to scale.

The concept of domain adaptation (DA) and domain generalization (DG) has been introduced to address the generalization problem and several solutions for skill learning have also been proposed. The main topic that encompasses both DA and DG is meta-learning [31, 113], where the proposition is to learn meta models (policies) and then adapt to new situations with minimum or no extra training. A possible future work could be to investigate meta-learning in the context of hybrid systems where additional dynamics modes and policy options (see Section 4.2) could be incrementally acquired, or examine the possibility of guaranteeing stability within a model-free meta RL.

Bibliography

- [1] R. Peter Bonasso, R. James Firby, E. Gat, D. Kortenkamp, D. P. Miller, and M. G. Slack, “Experiences with an architecture for intelligent, reactive agents,” *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 237–256, 1997.
- [2] K. B. Shimoga, “Robot grasp synthesis algorithms: A survey,” *The International Journal of Robotics Research*, vol. 15, no. 3, pp. 230–266, 1996.
- [3] S. M. LaValle, *Planning Algorithms*. Cambridge, U.K.: Cambridge University Press, 2006. Available at <http://planning.cs.uiuc.edu/>.
- [4] B. Siciliano and O. Khatib, *Springer handbook of robotics*. Springer, 2016.
- [5] B. J. Waibel and H. Kazerooni, “Theory and experiments on the stability of robot compliance control,” *IEEE Transactions on Robotics and Automation*, vol. 7, no. 1, pp. 95–104, 1991.
- [6] J. Park, *Control strategies for robots in contact*. PhD thesis, Stanford University, 2006.
- [7] B. Siciliano, L. Sciavicco, L. Villani, and G. Oriolo, *Robotics: modelling, planning and control*. Springer Science & Business Media, 2010.
- [8] S. Chiaverini, B. Siciliano, and L. Villani, “A survey of robot interaction control schemes with experimental comparison,” *IEEE/ASME Transactions on mechatronics*, vol. 4, no. 3, pp. 273–285, 1999.
- [9] N. Hogan and S. P. Buerger, “Impedance and interaction control,” in *Robotics and automation handbook*, pp. 375–398, CRC press, 2018.
- [10] E. Burdet, R. Osu, D. W. Franklin, T. E. Milner, and M. Kawato, “The central nervous system stabilizes unstable dynamics by learning optimal impedance,” *Nature*, vol. 414, no. 6862, pp. 446–449, 2001.
- [11] D. W. Franklin, E. Burdet, K. P. Tee, R. Osu, C.-M. Chew, T. E. Milner, and M. Kawato, “CNS learns stable, accurate, and efficient movements using a simple algorithm,” *Journal of neuroscience*, vol. 28, no. 44, pp. 11165–11173, 2008.

- [12] K. J. A. Kronander, *Control and Learning of Compliant Manipulation Skills*. PhD thesis, EPFL, 2015.
- [13] K. Kronander and A. Billard, “Online learning of varying stiffness through physical human-robot interaction,” in *2012 IEEE International Conference on Robotics and Automation*, pp. 1842–1849, 2012.
- [14] K. Kronander and A. Billard, “Learning compliant manipulation through kinesthetic and tactile human-robot interaction,” *IEEE transactions on haptics*, vol. 7, no. 3, pp. 367–380, 2013.
- [15] K. Kronander and A. Billard, “Passive interaction control with dynamical systems,” *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 106–113, 2015.
- [16] S. M. Khansari-Zadeh and O. Khatib, “Learning potential functions from human demonstrations with encapsulated dynamic and compliant behaviors,” *Autonomous Robots*, vol. 41, no. 1, pp. 45–69, 2017.
- [17] 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,” *Proceedings of Robotics: Science and Systems X (RSS 2014)*, vol. 10, p. 2014, 2014.
- [18] S. Levine, N. Wagener, and P. Abbeel, “Learning contact-rich manipulation skills with guided policy search,” in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 156–163, 2015.
- [19] M. Kalakrishnan, L. Righetti, P. Pastor, and S. Schaal, “Learning force control policies for compliant manipulation,” in *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pp. 4639–4644, 2011.
- [20] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, “Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks,” in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8943–8950, IEEE, 2019.
- [21] G. Thomas, M. Chien, A. Tamar, J. A. Ojea, and P. Abbeel, “Learning robotic assembly from CAD,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–9, 2018.
- [22] J. Luo, E. Solowjow, C. Wen, J. A. Ojea, A. M. Agogino, A. Tamar, and P. Abbeel, “Reinforcement learning on variable impedance controller for high-precision robotic assembly,” in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3080–3087, IEEE, 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 *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1010–1017, 2019.
- [24] M. Bogdanovic, M. Khadiv, and L. Righetti, “Learning variable impedance control for contact sensitive tasks,” *IEEE Robotics and Automation Letters*, 2020.
- [25] S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-end training of deep visuomotor policies,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1334–1373, 2016.
- [26] S. Gu, E. Holly, T. Lillicrap, and S. Levine, “Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates,” in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pp. 3389–3396, 2017.
- [27] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421–436, 2018.
- [28] M. P. Deisenroth, D. Fox, and C. E. Rasmussen, “Gaussian processes for data-efficient learning in robotics and control,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 2, pp. 408–423, 2015.
- [29] K. Chua, R. Calandra, R. McAllister, and S. Levine, “Deep reinforcement learning in a handful of trials using probabilistic dynamics models,” in *Advances in Neural Information Processing Systems 31* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), pp. 4754–4765, Curran Associates, Inc., 2018.
- [30] C. Finn, S. Levine, and P. Abbeel, “Guided cost learning: Deep inverse optimal control via policy optimization,” in *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*, (New York, New York, USA), pp. 49–58, PMLR, 20–22 Jun 2016.
- [31] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, “Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning,” in *Proceedings of the Conference on Robot Learning* (L. P. Kaelbling, D. Kragic, and K. Sugiura, eds.), vol. 100 of *Proceedings of Machine Learning Research*, pp. 1094–1100, PMLR, 30 Oct–01 Nov 2020.
- [32] S. A. Khader, H. Yin, P. Falco, and D. Kragic, “Stability-guaranteed reinforcement learning for contact-rich manipulation,” *IEEE Robotics and Automation Letters*, vol. 6, no. 1, pp. 1–8, 2020.

- [33] S. A. Khader, H. Yin, P. Falco, and D. Kragic, "Learning stable normalizing-flow control for robotic manipulation," *arXiv preprint arXiv:2011.00072*, 2020.
- [34] S. A. Khader, H. Yin, P. Falco, and D. Kragic, "Learning deep neural policies with stability guarantees," *arXiv preprint arXiv:2103.16432*, 2021.
- [35] S. A. Khader, H. Yin, P. Falco, and D. Kragic, "Data-efficient model learning and prediction for contact-rich manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4321–4328, 2020.
- [36] J.-J. E. Slotine, W. Li, *et al.*, *Applied nonlinear control*. Prentice hall Englewood Cliffs, NJ, 1991.
- [37] J. Lunze and F. Lamnabhi-Lagarigue, *Handbook of hybrid systems control: theory, tools, applications*. Cambridge University Press, 2009.
- [38] V. Gullapalli, R. A. Grupen, and A. G. Barto, "Learning reactive admittance control," in *Robotics and Automation, 1992. Proceedings., 1992 IEEE International Conference on*, pp. 1475–1480, 1992.
- [39] M. Nuttin and H. Van Brussel, "Learning the peg-into-hole assembly operation with a connectionist reinforcement technique," *Computers in Industry*, vol. 33, no. 1, pp. 101–109, 1997.
- [40] M. T. Mason, *Mechanics of robotic manipulation*. MIT press, 2001.
- [41] T. Lozano-Perez, M. T. Mason, and R. H. Taylor, "Automatic synthesis of fine-motion strategies for robots," *The International Journal of Robotics Research*, vol. 3, no. 1, pp. 3–24, 1984.
- [42] J. Hopcroft and G. Wilfong, "Motion of objects in contact," *The International journal of robotics research*, vol. 4, no. 4, pp. 32–46, 1986.
- [43] N. Hogan, "Impedance control: An approach to manipulation," in *American Control Conference, 1984*, pp. 304–313, IEEE, 1984.
- [44] M. Raibert and J. Craig, "Hybrid position/force control of manipulators1," *Journal of Dynamic Systems, Measurement, and Control*, vol. 102, p. 127, 1981.
- [45] O. Khatib, "A unified approach for motion and force control of robot manipulators: The operational space formulation," *IEEE Journal on Robotics and Automation*, vol. 3, no. 1, pp. 43–53, 1987.
- [46] J. K. Salisbury, "Active stiffness control of a manipulator in cartesian coordinates," in *1980 19th IEEE conference on decision and control including the symposium on adaptive processes*, pp. 95–100, IEEE, 1980.

- [47] C. Ott, R. Mukherjee, and Y. Nakamura, “Unified impedance and admittance control,” in *2010 IEEE International Conference on Robotics and Automation*, pp. 554–561, 2010.
- [48] J. Buchli, F. Stulp, E. Theodorou, and S. Schaal, “Learning variable impedance control,” *The International Journal of Robotics Research*, vol. 30, no. 7, pp. 820–833, 2011.
- [49] R. Sturges and S. Laowattana, “Fine motion planning through constraint network analysis,” in *Proceedings. IEEE International Symposium on Assembly and Task Planning*, pp. 160–170, IEEE, 1995.
- [50] A. J. Ijspeert, J. Nakanishi, and S. Schaal, “Learning attractor landscapes for learning motor primitives,” in *Advances in neural information processing systems*, pp. 1547–1554, 2003.
- [51] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, “Recent advances in robot learning from demonstration,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 297–330, 2020.
- [52] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [53] J. Kober, J. A. Bagnell, and J. Peters, “Reinforcement learning in robotics: A survey,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.
- [54] A. Bicchi, “Hands for dexterous manipulation and robust grasping: A difficult road toward simplicity,” *IEEE Transactions on robotics and automation*, vol. 16, no. 6, pp. 652–662, 2000.
- [55] S. Cruciani, B. Sundaralingam, K. Hang, V. Kumar, T. Hermans, and D. Kragic, “Benchmarking in-hand manipulation,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 588–595, 2020.
- [56] S. Schaal, “Dynamic movement primitives-a framework for motor control in humans and humanoid robotics,” in *Adaptive motion of animals and machines*, pp. 261–280, Springer, 2006.
- [57] A. Paraschos, C. Daniel, J. Peters, G. Neumann, *et al.*, “Probabilistic movement primitives,” *Advances in neural information processing systems*, 2013.
- [58] E. Gribovskaya, S. M. Khansari-Zadeh, and A. Billard, “Learning non-linear multivariate dynamics of motion in robotic manipulators,” *The International Journal of Robotics Research*, vol. 30, no. 1, pp. 80–117, 2011.
- [59] S. Calinon, F. Guenter, and A. Billard, “On learning, representing, and generalizing a task in a humanoid robot,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 2, pp. 286–298, 2007.

- [60] 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.
- [61] A. S. Polydoros and L. Nalpantidis, “Survey of model-based reinforcement learning: Applications on robotics,” *Journal of Intelligent & Robotic Systems*, vol. 86, pp. 153–173, May 2017.
- [62] J. Peters and S. Schaal, “Reinforcement learning of motor skills with policy gradients,” *Neural networks*, vol. 21, no. 4, pp. 682–697, 2008.
- [63] J. Peters, K. Mulling, and Y. Altun, “Relative entropy policy search,” in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.
- [64] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971*, 2015.
- [65] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *CoRR*, vol. abs/1707.06347, 2017.
- [66] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *International Conference on Machine Learning*, pp. 1861–1870, PMLR, 2018.
- [67] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, “Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7559–7566, 2018.
- [68] I. Lenz, R. A. Knepper, and A. Saxena, “DeepMPC: Learning deep latent features for model predictive control,” in *Robotics: Science and Systems*, 2015.
- [69] D. Nguyen-Tuong and J. Peters, “Model learning for robot control: a survey,” *Cognitive processing*, vol. 12, no. 4, pp. 319–340, 2011.
- [70] J. Rey, K. Kronander, F. Farshidian, J. Buchli, and A. Billard, “Learning motions from demonstrations and rewards with time-invariant dynamical systems based policies,” *Autonomous Robots*, vol. 42, no. 1, pp. 45–64, 2018.
- [71] A. Paraschos, E. Rueckert, J. Peters, and G. Neumann, “Probabilistic movement primitives under unknown system dynamics,” *Advanced Robotics*, vol. 32, no. 6, pp. 297–310, 2018.
- [72] A. Tamar, G. Thomas, T. Zhang, S. Levine, and P. Abbeel, “Learning from the hindsight plan—episodic MPC improvement,” in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pp. 336–343, 2017.

- [73] S. Hoppe, Z. Lou, D. Hennes, and M. Toussaint, “Planning approximate exploration trajectories for model-free reinforcement learning in contact-rich manipulation,” *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4042–4047, 2019.
- [74] J. Kober, K. Mülling, O. Krömer, C. H. Lampert, B. Schölkopf, and J. Peters, “Movement templates for learning of hitting and batting,” in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 853–858, 2010.
- [75] J. Kober and J. Peters, “Policy search for motor primitives in robotics,” *Machine Learning*, vol. 84, pp. 171–203, Jul 2011.
- [76] J. Kober, E. Oztop, and J. Peters, “Reinforcement learning to adjust robot movements to new situations,” in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.
- [77] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, “Safe model-based reinforcement learning with stability guarantees,” in *Advances in neural information processing systems*, pp. 908–918, 2017.
- [78] S. M. Richards, F. Berkenkamp, and A. Krause, “The lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems,” in *Conference on Robot Learning*, pp. 466–476, PMLR, 2018.
- [79] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, vol. 1. MIT press Cambridge, 2006.
- [80] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick, “End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks,” in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI*, pp. 3387–3395, 2019.
- [81] R. Cheng, A. Verma, G. Orosz, S. Chaudhuri, Y. Yue, and J. Burdick, “Control regularization for reduced variance reinforcement learning,” in *Proceedings of the 36th International Conference on Machine Learning, ICML*, pp. 1141–1150, 2019.
- [82] J. E. Colgate and N. Hogan, “Robust control of dynamically interacting systems,” *International journal of Control*, vol. 48, no. 1, pp. 65–88, 1988.
- [83] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, “Human-level control through deep reinforcement learning,” *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

- [84] M. P. Deisenroth, G. Neumann, J. Peters, *et al.*, “A survey on policy search for robotics,” *Foundations and Trends® in Robotics*, vol. 2, no. 1–2, pp. 1–142, 2013.
- [85] S. Kamthe and M. Deisenroth, “Data-efficient reinforcement learning with probabilistic model predictive control,” in *International Conference on Artificial Intelligence and Statistics*, pp. 1701–1710, PMLR, 2018.
- [86] S. Curi, F. Berkenkamp, and A. Krause, “Efficient model-based reinforcement learning through optimistic policy search and planning,” in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 14156–14170, Curran Associates, Inc., 2020.
- [87] J. Fu, S. Levine, and P. Abbeel, “One-shot learning of manipulation skills with online dynamics adaptation and neural network priors,” in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pp. 4019–4026, 2016.
- [88] R. Calandra, J. Peters, C. E. Rasmussen, and M. P. Deisenroth, “Manifold gaussian processes for regression,” in *Neural Networks (IJCNN), 2016 International Joint Conference on*, pp. 3338–3345, IEEE, 2016.
- [89] S. Paoletti, A. L. Juloski, G. Ferrari-Trecate, and R. Vidal, “Identification of hybrid systems a tutorial,” *European journal of control*, vol. 13, no. 2-3, pp. 242–260, 2007.
- [90] F. Lauer and G. Bloch, “Piecewise smooth system identification in reproducing kernel hilbert space,” in *53rd IEEE Conference on Decision and Control*, pp. 6498–6503, IEEE, 2014.
- [91] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts,” *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.
- [92] S. Linderman, M. Johnson, A. Miller, R. Adams, D. Blei, and L. Paninski, “Bayesian learning and inference in recurrent switching linear dynamical systems,” in *Artificial Intelligence and Statistics*, pp. 914–922, 2017.
- [93] P. Santana, S. Lane, E. Timmons, B. Williams, and C. Forster, “Learning hybrid models with guarded transitions,” in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
- [94] R. Y. Rubinstein and D. P. Kroese, *The Cross Entropy Method: A Unified Approach To Combinatorial Optimization, Monte-Carlo Simulation (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2004.
- [95] M. Wiering and M. Van Otterlo, “Reinforcement learning,” *Adaptation, learning, and optimization*, vol. 12, p. 3, 2012.

- [96] I. Kobyzev, S. Prince, and M. Brubaker, “Normalizing flows: An introduction and review of current methods,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [97] R. Ortega, A. J. Van Der Schaft, I. Mareels, and B. Maschke, “Putting energy back in control,” *IEEE Control Systems Magazine*, vol. 21, no. 2, pp. 18–33, 2001.
- [98] B. Amos, L. Xu, and J. Z. Kolter, “Input convex neural networks,” in *International Conference on Machine Learning*, pp. 146–155, 2017.
- [99] N. Hansen and A. Ostermeier, “Completely derandomized self-adaptation in evolution strategies,” *Evolutionary computation*, vol. 9, no. 2, pp. 159–195, 2001.
- [100] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber, “Natural evolution strategies,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 949–980, 2014.
- [101] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, “Evolution strategies as a scalable alternative to reinforcement learning,” *arXiv preprint arXiv:1703.03864*, 2017.
- [102] P. Florchinger, “Lyapunov-like techniques for stochastic stability,” *SIAM Journal on Control and optimization*, vol. 33, no. 4, pp. 1151–1169, 1995.
- [103] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” in *International conference on machine learning*, pp. 1889–1897, PMLR, 2015.
- [104] S. P. Buerger and N. Hogan, “Complementary stability and loop shaping for improved human–robot interaction,” *IEEE Transactions on Robotics*, vol. 23, no. 2, pp. 232–244, 2007.
- [105] S. M. Richards, F. Berkenkamp, and A. Krause, “The lyapunov neural network: Adaptive stability certification for safe learning of dynamical systems,” in *Proceedings of The 2nd Conference on Robot Learning*, vol. 87, pp. 466–476, PMLR, October 2018.
- [106] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, “A lyapunov-based approach to safe reinforcement learning,” in *Advances in Neural Information Processing Systems 31*, pp. 8092–8101, 2018.
- [107] A. A. Oliva, P. R. Giordano, and F. Chaumette, “A general visual-impedance framework for effectively combining vision and force sensing in feature space,” *IEEE Robotics and Automation Letters*, 2021.

- [108] E. F. Camacho and C. B. Alba, *Model predictive control*. Springer Science & Business Media, 2013.
- [109] P.-L. Bacon, J. Harb, and D. Precup, “The option-critic architecture,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.
- [110] A. Dutta, “Hybrid model based hierarchical reinforcement learning for contact rich manipulation task,” *KTH*, 2020.
- [111] K. Chatzilygeroudis, R. Rama, R. Kaushik, D. Goepp, V. Vassiliades, and J.-B. Mouret, “Black-box data-efficient policy search for robotics,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 51–58, 2017.
- [112] A. R. Teel, A. Subbaraman, and A. Sferlazza, “Stability analysis for stochastic hybrid systems: A survey,” *Automatica*, vol. 50, no. 10, pp. 2435–2456, 2014.
- [113] X. Chen, A. Ghadirzadeh, M. Björkman, and P. Jensfelt, “Adversarial feature training for generalizable robotic visuomotor control,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1142–1148, IEEE, 2020.

Part II

Included Publications

