An Algorithmic Perspective on Imitation Learning

Takayuki Osa
University of Tokyo
osa@edu.k.u-tokyo.ac.jp

Joni Pajarinen
Technische Universität Darmstadt
pajarinen@ias.tu-darmstadt.de

Gerhard Neumann
University of Lincoln
gneumann@lincoln.ac.uk

J. Andrew Bagnell
Carnegie Mellon University
dbagnell2@andrew.cmu.edu

Pieter Abbeel
University of California, Berkeley
pabbeel@cs.berkeley.edu

Jan Peters
Technische Universität Darmstadt
mail@jan-peters.net

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Robotics

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel and J. Peters. *An Algorithmic Perspective on Imitation Learning*. Foundations and Trends® in Robotics, vol. 7, no. 1-2, pp. 1–179, 2017.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-410-9

Foundations and Trends® in Robotics
Volume 7, Issue 1-2, 2017

Editorial Scope

Topics

Foundations and Trends® in Robotics publishes survey and tutorial articles in the following topics:

- Mathematical modelling
- Kinematics
- Dynamics
- Estimation methods
- Artificial intelligence in robotics
- Software systems and architectures
- Sensors and estimation
- Planning and control
- Human-robot interaction
- Industrial robotics
- Service robotics

Information for Librarians

Foundations and Trends® in Robotics, 2017, Volume 7, 4 issues. ISSN paper version 1935-8253. ISSN online version 1935-8261. Also available as a combined paper and online subscription.

Foundations and Trends® in Robotics
Vol. 7, No. 1-2 (2017) 1–179

DOI: 10.1561/2300000053

An Algorithmic Perspective on Imitation Learning

Takayuki Osa
University of Tokyo
osa@edu.k.u-tokyo.ac.jp

Joni Pajarinen
Technische Universität Darmstadt
pajarinen@ias.tu-darmstadt.de

Gerhard Neumann
University of Lincoln
gneumann@lincoln.ac.uk

J. Andrew Bagnell
Carnegie Mellon University
dbagnell2@andrew.cmu.edu

Pieter Abbeel
University of California, Berkeley
pabbeel@cs.berkeley.edu

Jan Peters
Technische Universität Darmstadt
mail@jan-peters.net

Contents

Abstract

As robots and other intelligent agents move from simple environments and problems to more complex, unstructured settings, manually programming their behavior has become increasingly challenging and expensive. Often, it is easier for a teacher to *demonstrate* a desired behavior rather than attempt to manually engineer it. This process of learning from demonstrations, and the study of algorithms to do so, is called *imitation learning*. This work provides an introduction to imitation learning. It covers the underlying assumptions, approaches, and how they relate; the rich set of algorithms developed to tackle the problem; and advice on effective tools and implementation.

We intend this paper to serve two audiences. First, we want to familiarize machine learning experts with the challenges of imitation learning, particularly those arising in robotics, and the interesting theoretical and practical distinctions between it and more familiar frameworks like statistical supervised learning theory and reinforcement learning. Second, we want to give roboticists and experts in applied artificial intelligence a broader appreciation for the frameworks and tools available for imitation learning.

We organize our work by dividing imitation learning into directly replicating desired behavior (sometimes called *behavioral cloning* [Bain and Sammut, 1996]) and learning the hidden objectives of the desired behavior from demonstrations (called *inverse optimal control* [Kalman, 1964] or *inverse reinforcement learning* [Russell, 1998]). In addition to method analysis, we discuss the design decisions a practitioner must make when selecting an imitation learning approach. Moreover, application examples—such as robots that play table tennis [Kober and Peters, 2009] and programs that play the game of Go [Silver et al., 2016]— illustrate the properties and motivations behind different forms of imitation learning. We conclude by presenting a set of open questions and point towards possible future research directions.

T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel and J. Peters. *An Algorithmic Perspective on Imitation Learning*. Foundations and Trends® in Robotics, vol. 7, no. 1-2, pp. 1–179, 2017.
DOI: 10.1561/2300000053.

1

Introduction

Programming autonomous behavior in machines and robots traditionally requires a specific set of skills and knowledge. However, human experts know how to demonstrate the desired task even if they do not know how to program the necessary behavior in a machine or robot. The purpose of imitation learning is to efficiently learn a desired behavior by imitating an expert's behavior. The application of imitation learning is not limited to physical systems. It can be a powerful tool to design autonomous behavior in systems such as web sites, computer games, and mobile applications. Any system that requires autonomous behavior similar to human experts can benefit from imitation learning.

However, imitation learning may be essential for robotics. It is now considered to be a key technology for applications such as manufacturing, elder care, and the service industry. These robots will be expected to work closely with humans in a dramatic shift from prior uses of robots. Powerful robotic manipulators are dangerous and have therefore been used mainly in constrained, predefined industrial applications; employees must undergo special training before working with them. This is changing due to recent advances in robotics from compute to the use of light, compliant, and safe robotic manipulators. They are ideal for applications where robots work alongside people, such as collaborating with human operators and reducing the physical workload of care givers. These applications require efficient, intuitive ways to teach robots the motions they need to perform from domain experts who may not possess special skills or knowledge about robotics.

In recent years, imitation learning has been investigated as a way to efficiently and intuitively program autonomous behavior[Schaal, 1999, Argall et al., 2009, Billard et al., 2008, Billard and Grollman, 2013, Bagnell, 2015, Billard et al., 2016]. In imitation learning, a human demonstrates how to perform a task. A robotic system learns a policy to execute the given task by imitating the demonstrated motions. Numerous imitation learning methods have been developed and imitation learning has become a gigantic field of research. As a consequence, capturing the entire field of imitation learning is not a trivial task.

The purpose of this survey is to provide a structural understanding of existing imitation learning methods and its relationship with other fields from supervised learning to control theory. We will describe what has been developed in the field of imitation learning and what should be developed in the future.

1.1 Key successes in Imitation Learning

One of the earliest and most well-known imitation learning success stories was the autonomous driving project Autonomous Land Vehicle In a Neural Network (ALVINN) at Carnegie Mellon University [Pomerleau, 1988]. In ALVINN, a neural network learned how to map input images to discrete actions in order to drive a vehicle. ALVINN's neural network had one hidden layer with five units. Its input layer had 30 by 32 units; its output layer had 30 units. Although the structure of this network was simple compared to modern neural networks with millions of parameters, the system succeeded at driving autonomously across the North American continent.

The Kendama robot developed by Miyamoto et al. [1996] is another successful application of imitation learning. In the early days of imitation learning, roboticists were mainly interested in teaching higher-level tasks from human demonstrations, such as "pick," "move," and "place" Kang and Ikeuchi [1993], Kuniyoshi et al. [1994]. In those settings, lower-level tasks were often considered to be simple, point-to-point motions. In the late 1990s, this focus shifted from task-level planning to trajectory-level planning. The term "learning from demonstration" has become very popular since its use by S. Schaal and G. Atkeson [Schaal, 1997, Atkeson and Schaal, 1997]. Since then, learning robot motions has been a key domain of imitation learning.

Recently, learning from human demonstrations has benefited from developments in deep neural networks. Recurrent neural networks such as long short-term memory (LSTM) networks Hochreiter and Schmidhuber [1997] have played a significant role in demonstrating how to succeed in many previously difficult sequential tasks by learning from demonstrated data. This includes tasks for generating handwriting Chung et al. [2015], natural language Wen et al. [2015], or image captions Karpathy and Fei-Fei [2015]. Furthermore, AlphaGo, the algorithm which was able to beat a human Go master and which we discuss in more detail in §3.4.2, initializes a deep neural network policy from human demonstrations Silver et al. [2016]. Often these recent approaches require a large amount of data. In §3, we will discuss how to learn from data to reproduce observed behavior in specific problem settings.

1.2 Imitation Learning from the Point of View of Robotics

Imitation learning is a class of methods that reproduces desired behavior based on expert demonstrations. In many cases, the experts are human operators and the learners are robotic systems, Thus, imitation learning is a technique that enables skills to be transferred from humans to robotic systems. To perform imitation learning, we need to develop a system that records demonstrations by experts and learns a policy to reproduce the demonstrated behavior from the recorded data. For this purpose, we need to answer the following questions.

General Aspects:

1. **Why and when should imitation learning be used?** This question clarifies the motivation for using imitation learning and what we should do with it.

2. **Who should demonstrate?** In many cases, the experts are human operators. Many imitation learning methods implicitly assume that demonstrations are provided by a single expert. When multiple experts are available, we need to decide which one should be imitated or how we can incorporate demonstrations from multiple experts.

3. **How should we record data of the expert demonstrations?** There are multiple ways of recording the behavior of experts. For example, motion capture systems and teleoperated robotic systems record data from expert behavior. This choice is closely related to the embodiment problem between experts and learners, which will be discussed in §3.7.1.

4. **What should we imitate?** The recorded data often includes redundant information about expert behavior. In such cases, features appropriate for the desired behavior should be selected. Meanwhile, the recorded data also includes unnecessary motions, which should not be imitated. The data must be segmented to extract the motions to be imitated.

Algorithmic Aspects:

5. **How should we represent the policy?** Expert behavior can be represented using methods such as symbolic representation, trajectory-based representation, and state-action space representation. The choice depends largely on the design of the entire system.

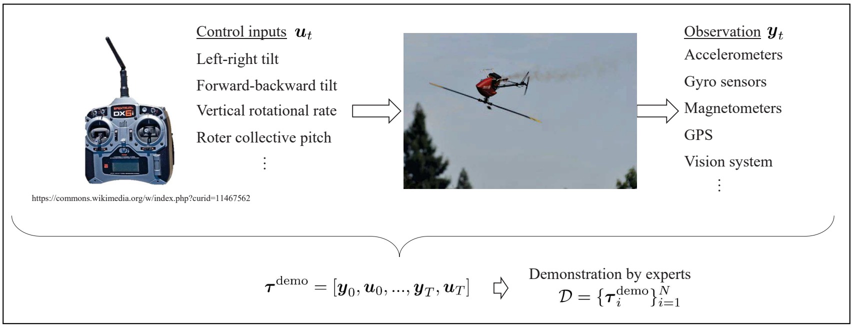
6. **How should we learn the policy?** Many algorithms for learning the policy have been developed over the past several decades. The choice of the algorithm for learning the policy is closely related to the choice of policy representation.

With regard to the first four questions, several survey papers on imitation learning [Argall et al., 2009, Billard et al., 2008, Billard and Grollman, 2013, Billard et al., 2016], provide a taxonomy of imitation learning from the perspective of robotics. Argall et al. [2009] indicate that it is essential to design an imitation learning system by considering the correspondence between the expert and the learner, data acquisition methods, and limitations of the demonstration dataset. Billard et al. [2008, 2016] provide an overview of imitation learning methods and highlight techniques for trajectory learning. However, none of the previous review articles focused on the *design decisions needed to develop new imitation learning algorithms* to enable answering questions five and six related to the algorithmic aspects discussed above. In addition, these articles did not discuss the algorithmic details of existing methods because the enormous amount of prior work on imitation learning makes it challenging to cover the entire range of previous studies.

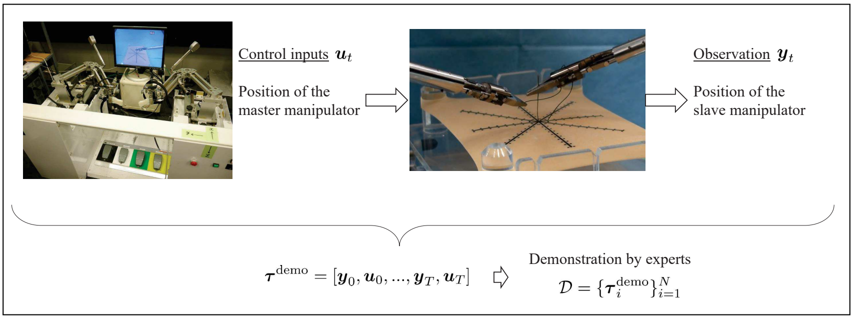
In this survey, we provide an overview of existing methods from the algorithmic point of view, which will be useful for both readers beginning the practice of imitation learning and readers who want to achieve a deeper understanding of the theoretical aspects of imitation learning. We discuss the design choices which one should consider in order to develop novel imitation learning algorithms. Although our survey cannot be exhaustive, we discuss the algorithmic details of existing algorithms as much as possible, which will be useful to readers who want to implement imitation learning techniques. Additionally, we develop an information theoretic understanding of existing methods, which will help readers to understand how existing methods relate to each other and figure out how they could be extended.

Let us illustrate how different design choices of imitation learning algorithms can be made in different applications. Figure 1.1 shows three applications of imitation learning: 1) an RC helicopter, 2) robotic surgery, and 3) quadruped robot locomotion. In these applications, design of the policies for motion planning and control vary. Abbeel et al. [2010] demonstrates acrobatic RC helicopter flight by learning from trajectories demonstrated by a human expert. In this system, the desired

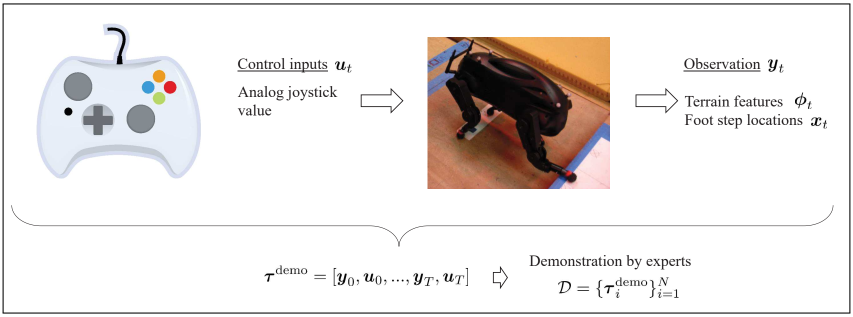
Full text available at: http://dx.doi.org/10.1561/2300000053



(a) Learning of acrobatic RC helicopter maneuvers [Abbeel et al., 2010]. The trajectories for acrobatic flights are learned from a human expert's demonstrations. To control the system with highly nonlinear dynamics, iterative learning control was used.



(b) Learning with a teleoperated system [Osa et al., 2014] where a position/velocity controller is available. To generalize the trajectory to different situations, a mapping from task situations to trajectories is learned from demonstrations under various situations.



(c) Learning quadruped robot locomotion [Zucker et al., 2011]. The footstep planning was addressed as an optimization of the reward/cost function, which was recovered from the expert demonstrations. Learning the reward/cost function allows the footstep planning strategy to be generalized to different terrains.

Figure 1.1: Observations $\boldsymbol{y}$ and control inputs $\boldsymbol{u}$ for imitation learning in (a) helicopter flight, (b) surgery, and (c) locomotion. Motion planning is formulated in different ways in these examples.

trajectories of acrobatic flights were learned from demonstrations with a supervised learning method. Osa et al. [2017b] also learned trajectories for autonomous knot tying from demonstrations by a human expert. To generalize a trajectory, Osa et al. [2017b] learned a direct mapping from task situations (contexts) to trajectories using demonstrations recorded under various situations. Contrary to [Abbeel et al., 2010, Osa et al., 2017b], Zucker et al. [2011] formulated footstep planning for quadruped robot locomotion as an optimization of the reward/cost function. The reward/cost function was recovered from demonstrations. In [Zucker et al., 2011], learning the reward/cost function as a function of terrain features enables the footstep planning strategy to be generalized to different terrains. Learning such reward/cost functions for manipulation tasks like as knot-tying [Osa et al., 2017b] is not trivial, since complex manipulation tasks often require nonlinear reward/cost functions.

Methods for learning policies also differ between applications. The observation and control inputs of the RC helicopter system are much noisier than those of the other two systems, and its dynamics are highly nonlinear [Abbeel et al., 2010]. Therefore, it is essential to estimate the true state using various sensory information and learn an adaptive controller through iterations of trials to achieve acrobatic RC helicopter flight. On the other hand, we can assume that the system state is precisely known and a position/velocity controller is available in the case of the tele-operation system in [Osa et al., 2014], which simplifies imitation learning significantly. In [Osa et al., 2014], the conditional trajectory distribution given a context can be learned with a simple regression method, and the planned trajectory can be executed by a standard velocity controller. In locomotion planning for a quadruped robot in [Zucker et al., 2011], estimating the reward/cost function requires an iterative learning process with virtual simulation of the learned policy. As one can see from these examples, learning methods can be very different between applications.

To apply imitation learning, it is essential to identify the structure of the system, formulate a given problem, and design an algorithm to solve the problem efficiently. In this survey, we focus on the algorithmic aspects of imitation and discuss necessary design choices, exploring various solutions proposed by previous studies.

In the rest of this chapter, we introduce several concepts in machine learning that are essential to understand imitation learning algorithms. We discuss the design choices of imitation learning algorithms in Chapter 2. We describe the details of behavioral cloning methods and inverse reinforcement learning methods in Chapters 3 and 4, respectively. To conclude, we list open questions of imitation learning in Chapter 5.

1.3 Key Differences between Imitation Learning and Supervised Learning

The imitation learning problem has special properties that distinguish it from the better known supervised learning setting [Shalev-Shwartz and Ben-David, 2014] : 1) the solution may have important structural properties including constraints (for example, robot joint limits), dynamic smoothness and stability, or leading to a coherent, multi-step plan [Bagnell, 2015]; 2) the interaction between the learner's decisions and its own input distribution (an *on-policy* versus *off-policy* distinction) , and 3) the increased necessity of minimizing the typically high cost of gathering examples.

As we learn a policy π from a dataset $\mathcal{D}$, imitation learning is closely related to supervised learning, and is particularly related to the field of *structured prediction* [Daumé III et al., 2009, Ratliff et al., 2006a, Taskar, 2005] , where the task is to learn a mapping from inputs $\boldsymbol{x}$ to a complex, structured output $\boldsymbol{y}$ (plans, parse trees, complex motions). Reductions of structured prediction to sequential decision [Daumé III et al., 2009], and reductions of imitation learning to structured prediction [Ratliff et al., 2006b] show the close connection, and cross-fertilization between these research areas has been important for both. In practice, distinctions arise because of the structural properties of policies we attempt to imitate, and the difficulty of "resetting" state and restarting predictions is too costly or even infeasible in most imitation learning settings because a physical system is often involved.

In addition, it is often the case that the embodiments of the expert and the learner are different. For example, when transferring human skills to a humanoid robot, the motion captured from a human expert may be infeasible for the humanoid. In such a case, the demonstrated motion needs to be adapted to be feasible for the humanoid. This kind of adaptation is less common in the standard supervised learning.

In machine learning, the prediction problem where the source domain distribution and the target domain distribution are different is often referred to as "*covariate shift*" or "*domain adaptation*" [Sugiyama, 2015]. In imitation learning, the source domain corresponds to expert demonstrations and the target domain to learner reproductions. In imitation learning, the demonstration dataset does not cover all possible situations since collecting expert demonstrations to cover all situations is usually too expensive and time-consuming. As a result, the learner often encounters states which were not encountered by the expert during demonstrations, which means that the target domain distribution is different from the source distribution. Therefore, covariate shift or domain adaptation is closely related to imitation learning [Bagnell, 2015].

Imitation learning is also closely related to reinforcement learning (RL), which tries to obtain a policy that maximizes an expected reward [Sutton and Barto, 1998] signal. In RL, we employ a reward function that encourages a desired behavior. However, in imitation learning we often assume optimal (or at least "good") expert demonstrations which are not available in basic reinforcement learning, and which provide prior knowledge that allows for dramatically more efficient methods. Recent work by Sun et al. [2017] demonstrates a potentially *exponential* decrease in sample complexity in learning a task by imitation rather than by trial-and-error reinforcement learning, and empirical results have long shown such benefits [Silver et al., 2016, Kober and Peters, 2009, Abbeel et al., 2010]. Moreover, in the imitation learning setting, as we detail below, we may or may not have access to a true reward function.

1.4 Insights for Machine Learning and Robotics Research

As imitation learning offers intuitive ways to program robotic motions by demonstrating the desired motion, imitation learning attracted interests from robotic researchers. The robotics community has developed many imitation learning methods for motion planning and robot control. When planning a trajectory for a robotic system, it is often necessary to make sure that a planned trajectory satisfies some constraints such as smooth convergence to a new goal state. For this reason, robotics researchers have developed "custom" trajectory representations that explicitly satisfy constraints necessary for robotic applications. Machine learning techniques are often used as a part of such frameworks. However, robotics researchers need to be aware that rich set of algorithms have been developed by the machine learning community and some of new algorithms might eliminate the need for customizing policy or trajectory representation.

For machine learning researchers, imitation learning offers interesting practical and theoretical problems, which differ from standard supervised and reinforcement learning settings. Although imitation learning is closely related to structured prediction, it is often challenging to apply existing machine learning methods to imitation learning, especially robotic applications. In imitation learning, collecting demonstrations and performing rollouts are often expensive and time-consuming. Therefore, it is necessary to consider how to minimize these costs and perform learning efficiently. In addition, embodiments and observability of the learner and the expert are different in many applications. In such cases, the demonstrated motion needs to be adapted based on the learner's embodiment and observability. These difficulties in imitation learning present new challenges to machine learning researchers.

1.5 Statistical Machine Learning Background

To understand imitation learning algorithms, familiarity with several concepts in statistical machine learning is essential. In this section, we briefly introduce the notation we use and these concepts.

1.5.1 Notation and Mathematical Formalization

Before introducing important concepts in machine learning, we introduce the notation in this article. Table 1.1 summarizes our notation. Throughout this survey, we use the bold style for vector values, and the non-bold style for scalar values. Demonstrations by an expert are often given as a set of trajectories. In this case, the dataset of demonstrations is given by $\mathcal{D} = \{\boldsymbol{\tau}^0, \ldots, \boldsymbol{\tau}^m\}$. We use the lower script to denote the time index; $\boldsymbol{x}_t$ represents the state of the system at time step t. We review many methods that manipulate probability distributions in various ways. To make equations concise, the probability distribution induced by the experts' policy is denoted by q, and the distribution induced by the learner's policy is denoted by p. For example, $p(\boldsymbol{\tau})$ represents the probability distribution over trajectories induced by the learner's policy. The term "action" is mainly used in machine learning community, and "control input" is mainly used in robotic community and control theory community. Since imitation learning methods have been developed in all of these communities, we use the word "action"

Table 1.1: Table of Notation. We use a notation common in the control literature for states and controls.

$\boldsymbol{x}$	system state
$\boldsymbol{s}$	context
$\boldsymbol{\phi}$	feature vector
$\boldsymbol{u}$	control input/action
$\boldsymbol{\tau}$	trajectory
π	policy
$\mathcal{D}$	dataset of demonstrations
q	probability distribution induced by an expert's policy
p	probability distribution induced by a learner's policy
t	time
T	finite horizon
N	number of demonstrations
E	superscript representing an expert e.g. π^{E} denotes an expert's policy
L	superscript representing a learner e.g. π^{L} denotes a learner's policy
demo	superscript representing a demonstration by an expert e.g. $\boldsymbol{\tau}^{\mathrm{demo}}$ denotes a trajectory demonstrated by an expert

and "control input" interchangeably. We use the term "context" to refer to the condition relevant to the task. The context $\boldsymbol{s}$ can be the initial state of the system $\boldsymbol{x}_0$ or the state of relevant objects. For instance, the position of the ball can be part of the context in a hitting-a-ball task. We use T to denote the finite horizon of the trajectory. Therefore, the total number of the time steps of a single trajectory is $T+1$ in our notation.

1.5.2 Markov Property

A sequence of states $\boldsymbol{x}_0, ..., \boldsymbol{x}_t$ is a Markov chain if at any time t, the future states $\boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+2}, ...$ depend on the history $\boldsymbol{x}_0, ..., \boldsymbol{x}_t$ only through the present state $\boldsymbol{x}_t$ [Serfozo, 2009]. In other words, the next state $\boldsymbol{x}_{t+1}$ only depends on the current state $\boldsymbol{x}_t$ in a Markov chain. This property is called the *Markov property*.

1.5.3 Markov Decision Process

A Markov decision process (MDP) is a process that satisfies the Markov property. If the state and action spaces are finite, then it is called a finite Markov decision process (finite MDP) [Sutton and Barto, 1998]. An MDP is defined as a tuple $(\mathcal{X}, \mathcal{U}, \mathcal{P}, \gamma, D, R)$. $\mathcal{X}$ is a finite set of states; $\mathcal{U}$ is a set of control inputs; $\mathcal{P}$ is a set of state transitions probabilities; $\gamma \in [1, 0)$ is a discount factor; D is the initial-state distribution from which the initial state $\boldsymbol{x}_0$ is drawn; and $R : \mathcal{X} \mapsto \mathbb{R}$ is the reward function.

1.5.4 Entropy

Given the random variable $\boldsymbol{x}$ and its probability distribution $p(\boldsymbol{x})$, the entropy

$$H(p) = -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} \tag{1.1}$$

is defined as the amount of information conveyed by transmitting $\boldsymbol{x}$ [Bishop, 2006]. Note that the entropy $H(\boldsymbol{x})$ is a convex function.

1.5.5 Kullback-Leibler (KL) Divergence

In the field of information geometry, the KL divergence is used to quantify a difference between two probability distributions[Kullback and Leibler, 1951], i.e.,

$$D_{\mathrm{KL}}\left(p(\boldsymbol{x})||q(\boldsymbol{x})\right) = \int p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}. \tag{1.2}$$

Since the KL divergence identifies a difference between two probability distributions, it is useful for cases in which stochastic policies are going to be learned, or stochastic trajectories result from a deterministic policy. Please note that the KL divergence is not symmetric, therefore $D_{\mathrm{KL}}\left(p||q\right) \neq D_{\mathrm{KL}}\left(q||p\right)$. The KL divergence can be obtained as a Bregman divergence derived from the negative entropy [Amari, 2016] and is widely used as a measure in multiple imitation learning approaches.

1.5.6 Information and Moment Projections

One common approach to learning a policy from a dataset is to consider "projecting" that dataset onto the space of the policy model. Information theory emphasizes two kinds of projections: the Information(I)-projection and the Moment(M)-projection [Bishop, 2006]. Using the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951], the I-projection is

$$p^* = \arg\min_p D_{\mathrm{KL}}(p \parallel q) \ , \tag{1.3}$$

and, the M-projection

$$p^* = \arg\min_p D_{\mathrm{KL}}(q \parallel p) \ . \tag{1.4}$$

As the KL divergence is not symmetric, these two projections result in different solutions when a given distribution is multi-modal as shown in Figure 1.2. While the M-projection averages over the several modes, the I-projection concentrates on a single mode. Performing the I-projection is often not straight-forward, although the M-projection can often be performed relatively easily by maximizing the likelihood with respect to a given training dataset [Bishop, 2006].

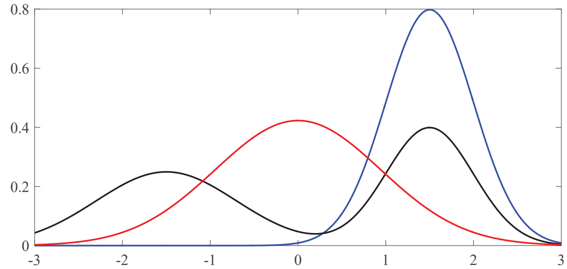


Figure 1.2: Illustration of I- and M- projections. Given a distribution with two modes as shown in black, M-projection will give a solution that averages over two modes as shown in red. On the contrary, I-projection will give a solution that concentrates on one of the modes.

1.5.7 The Maximum Entropy Principle

Let us consider a probability distribution $p(\boldsymbol{x})$ that matches the features of an unknown distribution q, i.e. it satisfies

$$\mathbb{E}_p[\boldsymbol{\phi}(\boldsymbol{x})] = \mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{x})],$$

where $q(\boldsymbol{x})$ is an unknown probability distribution and $\mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{x})]$, which is the expectation of a feature function $\boldsymbol{\phi}(\boldsymbol{x})$, is available. As there are typically an infinite amount of such distributions, we need an additional constraint to obtain a unique solution [Amari, 2016].

The maximum entropy principle [Jaynes, 1957] suggests to choose a distribution that maximizes the entropy

$$H(p) = -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x}$$

among the distributions that satisfy $\mathbb{E}_p[\boldsymbol{\phi}(\boldsymbol{x})] = \mathbb{E}_q[\boldsymbol{\phi}(\boldsymbol{x})]$. From this constrained optimization program, the maximum entropy distribution can be computed as

$$p(\boldsymbol{x}) \propto \exp\left(\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})\right), \tag{1.5}$$

where $\boldsymbol{w}$ is a vector-valued Lagrangian multiplier for the feature matching constraint. While the maximum entropy principle does not directly translate into a practical algorithm, it uncovers an interesting observation. Every distribution that is in a log-linear representation given by Equation 1.5, is the maximum entropy distribution that can match specific feature expectations given by the feature vector $\boldsymbol{\phi}(\boldsymbol{x})$. This is true for typical distributions from the exponential family such as the Gaussian distribution, which is the maximum entropy distribution that matches first and second order moments. The notion of Maximum Entropy generalizes to Maximum Causal Entropy, which turns out to be a natural notion of uncertainty for dynamical systems [Ziebart et al., 2013].

1.5.8 Background: Reinforcement Learning

Reinforcement learning is a class of methods that autonomously learns policies through iterations of trials and evaluations. The goal of reinforcement learning is to learn a policy π that maps the state of the system to the control input so as to maximize the expected reward $J(\pi)$. The reward r_t represents the quality of the given state, action or trajectory at time t. For example, r_t could be large when a robot is close to the desired trajectory and small when the robot is far from the trajectory, or, r_t could be large for stable robot grasps and small for unstable ones. With a finite horizon T, the expected return is given by the accumulation of the reward at each time step,

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} r_t \middle| \pi \right]. \tag{1.6}$$

Alternatively, the discounted accumulated reward is used for the infinite horizon scenario, i.e.,

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| \pi \right], \tag{1.7}$$

where the discounted factor γ controls the trade-off between shorter term rewards and longer term rewards. The desired policy π^* is given by

$$\pi^* = \arg\max_{\pi} J(\pi). \tag{1.8}$$

The value of a state $\boldsymbol{x}$ under a policy π can be computed as the expected reward when starting from $\boldsymbol{x}$ and following π

$$V^{\pi}(\boldsymbol{x}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| \boldsymbol{x}_0 = \boldsymbol{x}, \pi \right]. \tag{1.9}$$

$V^\pi(\boldsymbol{x}_t)$ is often called the *value function* [Sutton and Barto, 1998]. Likewise, the value of taking action $\boldsymbol{u}$ in state $\boldsymbol{x}$ under a policy π can be computed as the expected reward when starting from the action $\boldsymbol{u}$ in a state $\boldsymbol{x}$ and thereafter following policy π

$$Q^\pi(\boldsymbol{x}, \boldsymbol{u}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| \boldsymbol{x}_0 = \boldsymbol{x}, \boldsymbol{u}_0 = \boldsymbol{u}, \pi\right]. \tag{1.10}$$

$Q^\pi(\boldsymbol{x}_t, \boldsymbol{u}_t)$ is often called the *action-value function* [Sutton and Barto, 1998].

For an overview of reinforcement learning methods, please refer to [Sutton and Barto, 1998, Szepesvari, 2010, Wiering and van Otterlo, 2012, Sugiyama et al., 2013] and for an overview in reinforcement learning in robotics, please refer to Kober et al. [2013], Deisenroth et al. [2013b].

1.6 Formulation of the Imitation Learning Problem

The goal of imitation learning is to learn a policy that reproduces the behavior of experts who demonstrate how to perform the desired task. Suppose that the behavior of the expert demonstrator (or the learner itself) can be observed as a trajectory $\boldsymbol{\tau} = [\boldsymbol{\phi}_0, ..., \boldsymbol{\phi}_T]$, which is a sequence of features $\boldsymbol{\phi}$. The features $\boldsymbol{\phi}$, which can be the state of the robotic system or any other measurements, can be chosen according to the given problem. Please note that the features ϕ do not have to be manually specified, and $\boldsymbol{\phi}$ could be as general as simply pixels in raw images.

Often, the demonstrations are recorded under different conditions, for example, grasping an object at different locations. We will refer to these task conditions as context vector $\boldsymbol{s}$ of the task which is stored together with the feature trajectories. The context $\boldsymbol{s}$ can contain any information relevant to the task, e.g., the initial state of the robotic system or positions of target objects. Note that, as the context describes the current task, it is typically fixed during task execution and the only dynamic aspects of the problem are the state features $\boldsymbol{\phi}_t$. Optionally, a reward signal r that the expert is trying to optimize is also available in some problem settings [Ross and Bagnell, 2014].

In imitation learning, we collect a dataset of demonstrations $\mathcal{D} = \{(\boldsymbol{\tau}_i, \boldsymbol{s}_i, r_i)\}_{i=1}^{N}$ that consists of pairs of trajectories $\boldsymbol{\tau}$, contexts $\boldsymbol{s}$, and optionally reward signals r. The data collection process can be both offline and online. Using the collected dataset $\mathcal{D}$, a common *optimization-based strategy* learns a policy π^* that satisfies

$$\pi^* = \arg\min D\left(q(\boldsymbol{\phi}), p(\boldsymbol{\phi})\right), \tag{1.11}$$

where $q(\boldsymbol{\phi})$ is the distribution of the features induced by the experts' policy, $p(\boldsymbol{\phi})$ is the distribution of the features induced by the learner, and $D(q, p)$ is a similarity measure between q and p. Both offline and online learning scenarios of this problem have been considered [Ross et al., 2011]. Please note that, when the dataset contains demonstrations of multiple tasks and the contexts include information of each task, this problem can be considered multitask learning as in recent work by Duan et al. [2017], Finn et al. [2017a,b].

In addition, we often have access to an environment such as a simulator or a physical robotic system where we can perform and evaluate a policy through interaction. This simulator can be used to gather new data and iteratively improve the policy to better match the demonstrations.

References

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the international conference on Machine learning (ICML)*, 2004.

P. Abbeel, A. Coates, and A. Y. Ng. Autonomous helicopter aerobatics through apprenticeship learning. *The International Journal of Robotics Research*, 29(13):1608–1639, 2010.

S. Amari. *Information Geometry and Its Applications*. Springer, 2016.

H. Ben Amor, G. Neumann, S. Kamthe, O. Kroemer, and J. Peters. Interaction primitives for human-robot cooperation tasks. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2831–2837, 2014.

B. Anderson and J Moore. *Optimal Control: Linear Quadratic Methods*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1990.

O. Arenz, H. Abdulsamad, and G. Neumann. Optimal control and inverse optimal control by distribution matching. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5): 469–483, 2009.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 214–223, 2017.

C. G. Atkeson and S. Schaal. Robot learning from demonstration. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12–20, 1997.

C. G. Atkeson, Andrew W. Moore, and Stefan Schaal. Locally weighted learning for control. *Artificial Intelligence Review*, 11(1):75–113, 1997. ISSN 1573-7462. . URL `http://dx.doi.org/10.1023/A:1006511328852`.

J. Andrew (Drew) Bagnell. An invitation to imitation. Technical report, Robotics Institute, Carnegie Mellon University, March 2015.

M. Bain and C. Sammut. A framework for behavioural cloning. *Machine Intelligence 15*, pages 103–129, 1996.

C. L. Baker, R. Saxe, and J. B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

G. BakIr, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S.V.N Vishwanathan. *Predicting Structured Data (Neural Information Processing)*. MIT Press, 2007.

N. Baram, O. Anschel, I. Caspi, and S. Mannor. End-to-end differentiable adversarial imitation learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

A. Billard and D.H. Grollman. Learning by demonstration. *Scholarpedia*, 2013. .

A. Billard, S. Calinon, R. Dillmann, and S. Schaal. *Springer handbook of robotics*, chapter Robot programming by demonstration, pages 1371–1394. Springer Berlin Heidelberg, 2008.

A. Billard, S. Calinon, and R. Dillmann. *Handbook of robotics*, chapter Learning from Humans, pages 1995–2014. Springer, 2016.

C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

M. Bloem and N. Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *Proceedings of the IEEE Conference on Decision and Control (CDC)*, pages 4911–4916, 2014.

K. Bogert and P. Doshi. Multi-robot inverse reinforcement learning under occlusion with interactions. In *Proceedings of the International Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, pages 173–180, 2014.

K. Bogert and P. Doshi. Toward estimating others transition models under occlusion for multi-robot irl. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1867–1873, 2015.

K. Bogert, J. F. Lin, P. Doshi, and D. Kulic. Expectation-maximization for inverse reinforcement learning with hidden data. In *Proceedings of the International Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, pages 1034–1042, 2016.

A. Boularias, J. Kober, and J Peters. Relative entropy inverse reinforcement learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTAT)*, 2011.

A. Boularias, O. Krömer, and J. Peters. Structured apprenticeship learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, pages 227–242. Springer, 2012.

D. M. Bradley. *Learning in modular systems*. PhD thesis, Carnegie Mellon University, 2010.

S. Calinon. Robot learning with task-parameterized generative models. In *In Proceedings of International Symposium on Robotics Research (ISRR)*, 2015.

S. Calinon. A tutorial on task-parameterized movement learning and retrieval. *Intelligent Service Robotics (Springer)*, 9(1):1–29, 2016.

S. Calinon and A. Billard. Incremental learning of gestures by imitation in a humanoid robot. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 255–262, 2007.

S. Calinon and A. Billard. Statistical learning by imitation of competing constraints in joint space and task space. *Advanced Robotics*, 23(15):2059–2076, 2009.

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*, 37(2):286–298, 2007.

S. Calinon, F. D'halluin, E. L. Sauser, D. G. Caldwell, and A. G Billard. Learning and reproduction of gestures by imitation. *IEEE Robotics & Automation Magazine*, 17(2):44–54, 2010.

S. Calinon, A. Pistillo, and D. G. Caldwell. Encoding the time and space constraints of a task in explicit-duration hidden Markov model. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3413–3418. IEEE, 2011.

R. Camacho and D. Michie. Behavioral cloning: A correction. *AI MAGAZINE*, 16(2), 1995.

S. Cambon, R. Alami, and Fabien Gravot. A hybrid approach to intricate motion, manipulation and task planning. *The International Journal of Robotics Research*, 28:104–126, 2009.

R. A. Chambers and D. Michie. Man-machine co-operation on a learning task. *Computer Graphics: Techniques and Applications*, 1969.

K. Chang, A. Krishnamurthy, A. Agarwal, H. Daumé III, and J. Langford. Learning to search better than your teacher. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015. URL `http://hal3.name/docs/#daume15lols`.

S. Chernova and M. Veloso. Interactive policy learning through confidence-based autonomy. *Journal of Artificial Intelligence Research*, 34:1–25, 2009.

J. Choi and K. Kim. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, 12(Mar):691–730, 2011a.

J. Choi and K. E. Kim. Map inference for bayesian inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1989–1997, 2011b.

J. Choi and K. E. Kim. Hierarchical bayesian inverse reinforcement learning. *IEEE Transactions on Cybernetics*, 45(4):793–805, April 2015. ISSN 2168-2267. .

H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2):114–141, 2003.

J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2980–2988, 2015.

A. Coates, P. Abbeel, and A. Y. Ng. Learning for control from multiple demonstrations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, 1995.

C. Daniel, O. Kroemer, M. Viering, J. Metz, and J. Peters. Active reward learning with a novel acquisition function. *Autonomous Robots*, 39(3):389–405, 2015.

H. Daumé III and J. Langford. Advances in structured prediction. In *Tutorials in the International Conference on Machine Learning (ICML)*, July 2015.

H. Daumé III, J. Langford, and D. Marcu. Search-based structured prediction. *Machine Learning*, 75:297–325, 2009.

R. Dearden, N. Friedman, and D. Andre. Model based bayesian exploration. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 150–159. Morgan Kaufmann Publishers Inc., 1999.

M. P. Deisenroth. Efficient reinforcement learning using gaussian processes. *KIT Scientific Publishing*, 2010.

M. P. Deisenroth and C. E. Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

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*, 37(2):408–423, 2013a.

M. P. Deisenroth, G. Neumann, and J. Peters. A survey on policy search for robotics. *Foundations and Trends in Robotics*, 2(1-2):1–142, 2013b.

M. P. Deisenroth, P. Englert, J. Peters, and D. Fox. Multi-task policy search for robotics. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3876–3881, 2014.

M. Deniša, A. Gams, A. Ude, and T. Petrič. Learning compliant movement primitives through demonstration and statistical generalization. *IEEE/ASME Transactions on Mechatronics*, 21(5):2581–2594, 2016.

Andreas Doerr, Nathan Ratliff, Jeannette Bohg, Marc Toussaint, and Stefan Schaal. Direct loss minimization inverse optimal control. *Proceedings of Robotics: Science and Systems (R:SS)*, pages 1–9, 2015.

F. Doshi-Velez, J. Pineau, and N. Roy. Reinforcement learning with limited reinforcement: Using bayes risk for active learning in pomdps. *Artificial Intelligence*, 187:115–132, 2012.

F. Doshi-Velez, D. Pfau, F. Wood, and N. Roy. Bayesian nonparametric methods for partially-observable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):394–407, 2015.

A. D. Dragan, K. Muelling, J. Andrew Bagnell, and S. S. Srinivasa. Movement primitives via optimization. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2339–2346, May 2015. .

Y. Duan, M. Andrychowicz, B. C. Stadie, J. Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba. One-shot imitation learning. *arXiv preprint*, abs/1703.07326, 2017. URL `http://arxiv.org/abs/1703.07326`.

M. Dudík and R. E. Schapire. Maximum entropy distribution estimation with generalized regularization. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, pages 123–138, 2006.

K. Dvijotham and E. Todorov. Inverse optimal control with linearly-solvable mdps. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

S. Ekvall and D. Kragic. Robot learning from demonstration: a task-level planning approach. *International Journal of Advanced Robotic Systems*, 5 (3), 2008.

P. Englert, A. Paraschos, J. Peters, and M. P. Deisenroth. Probabilistic model-based imitation learning. *Adaptive Behavior*, 21:388–403, 2013.

M. Ewerton, G. Neumann, R. Lioutikov, H. Ben Amor, J. Peters, and G. Maeda. Learning multiple collaborative tasks with a mixture of interaction primitives. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1535–1542, 2015.

M. Ewerton, G. Maeda, G. Kollegger, J. Wiemeyer, and J. Peters. Incremental imitation learning of context-dependent motor skills. In *Proceedings of IEEE international Conference on Humanoid Robots (HUMANOIDS)*, 2016.

P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B*, 69(4):507–740, 2007.

C. Finn, P. Christiano, P. Abbeel, and S. Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. In *arXiv 1611.03852*, 2016a.

C. Finn, S. Levine, and P. Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016b.

C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning (ICML),*, 2017a.

C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine. One-shot visual imitation learning via meta-learning. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2017b.

E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Sharing features among dynamical systems with beta processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

A. Gams, B. Nemec, A. J. Ijspeert, and A. Ude. Coupling movement primitives: Interaction with the environment and bimanual tasks. *IEEE Transactions on Robotics*, 30(4):816–830, Aug 2014. ISSN 1552-3098. .

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

A. Graves, A.-R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASPP)*, pages 6645–6649. IEEE, 2013.

E. Gribovskaya, S. M. Khansari-Zadeh, and A. Billard. Learning non-linear multivariate dynamics of motion in robotic manipulators. *International Journal of Robotics Research*, 30(1):80–117, 2011.

D. B. Grimes and R. P. N. Rao. *Creating brain-like intelligence: From basic principles to complex intelligent systems*, chapter Learning Actions through Imitation and Exploration: Towards Humanoid Robots That Learn from Humans, pages 103–138. Springer Berlin Heidelberg, 2009.

D. B. Grimes, R. Chalodhorn, and R. P. N. Rao. Dynamic imitation in a humanoid robot through nonparametric probabilistic inference. In *Proceedings of Robotics: Science and Systems (R:SS)*, 2006a.

D. B. Grimes, D. R. Rashid, and R. P. Rao. Learning nonparametric models for probabilistic imitation. In *Advances in Neural Information Processing Systems 19*, 2006b.

A. Grubb and J. A. Bagnell. Boosted backpropagation learning for training deep modular networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

A. Gupta, C. Devin, Y. Liu, P. Abbeel, and S. Levine. Learning invariant feature spaces to transfer skills with reinforcement learning. In *Proceedings of the Internatinal Conference on Learning Representatinos (ICLR)*, 2017.

D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

M. Haruno, D.M. Wolpert, and M. Kawato. Mosaic model for sensorimotor learning and control. *Neural Computation*, 13(10):2201–2220, 2001.

I. Havoutis and S. Calinon. Supervisory teleoperation with online learning and optimal control. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

P. Henderson, W. Chang, P. L. Bacon, D. Meger, J. Pineau, and D. Precup. Optiongan: Learning joint reward-policy options using generative adversarial inverse reinforcement learning. In *In the Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

J. Ho and S. Ermon. Generative adversarial imitation learning. In *Adances in Neural Information Processing Systems (NIPS)*, 2016.

J. Ho, J. K. Gupta, and S. Ermon. Model-free imitation learning with policy optimization. In *Proceedings of the International Conference on International Conference on Machine Learning (ICML)*, 2016.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

H. Hoffmann, P. Pastor, D. H. Park, and S. Schaal. Biologically-inspired dynamical systems for movement generation: Automatic real-time goal adaptation and obstacle avoidance. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 2587–2592, 2009.

S. Huang, J. Pan, G. Mulcaire, and P. Abbeel. Leveraging appearance priors in non-rigid registration with applications to manipulation of deformable objects. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

A. J. Ijspeert, J. Nakanishi, and S. Schaal. Learning attractor landscapes for learning motor primitives. In *Advances in Neural Information Processing Systems (NIPS)*, 2002a.

A. J. Ijspeert, J. Nakanishi, and S. Schaal. Movement imitation with nonlinear dynamical systems in humanoid robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1398–1403, 2002b.

A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal. Dynamical movement primitives: Learning attractor models for motor behaviors. *Neural Computation*, 25(2):328–373, 2013.

T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura. Embodied symbol emergence based on mimesis theory. *The International Journal of Robotics Research*, 2004.

R. A. Jacobs, M. I. Jordan, S. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 1991.

A. Jain, S. Sharma, T. Joachims, and A. Saxena. Learning preferences for manipulation tasks from online coactive feedback. *The International Journal of Robotics Research*, 2015.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, May 1957. . URL `http://link.aps.org/doi/10.1103/PhysRev.106.620`.

M. Kalakrishnan, P. Pastor, L. Righetti, and S. Schaal. Learning objective functions for manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1331–1336. IEEE, 2013.

R. E. Kalman. When is a linear control system optimal? *Trans. ASME, J. Basic Eng., Ser. D.*, 86(1):51 – 60, 1964.

S. B. Kang and K. Ikeuchi. Toward automatic robot instruction from perception-recognizing a grasp from observation. *IEEE Transactions on Robotics and Automation*, 9(4):432–443, Aug 1993.

A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.

H. Khalil. *Nonlinear systems.* Prentice Hall, Upper Saddle River, NJ, 1996.

S. M. Khansari-Zadeh and A. Billard. Learning stable nonlinear dynamical systems with gaussian mixture models. *IEEE Transactions on Robotics*, 27 (5):943–957, 2011.

S. M. Khansari-Zadeh and A. Billard. Learning control lyapunov function to ensure stability of dynamical system-based robot reaching motions. *Robotics and Autonomous Systems*, 62(6):752–765, 2014.

S. Kim, A. Shukla, and A. Billard. Catching objects in flight. *IEEE Transactions on Robotics*, 30(5):1049–1065, 2014.

K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *European Conference on Computer Vision (ECCV)*, pages 201–214. Springer, 2012.

J. Kober and J. Peters. Learning motor primitives for robotics. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 2112–2118, 2009.

J. Kober, B. Mohler, and J. Peters. Learning perceptual coupling for motor primitives. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robot Systems (IROS)*, pages 834–839, 2008.

J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32:1238–1274, 2013.

J. Kohlmorgen and S. Lemm. A dynamic hmm for on-line segmentation of sequential data. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, pages 793–800, 2001.

J. Z. Kolter, P. Abbeel, and A. Y. Ng. Hierarchical apprenticeship learning with application to quadruped locomotion. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

G. Konidaris, S. Kuindersma, R. Grupen, and A. Barto. Robot learning from demonstration by constructing skill trees. *The International Journal of Robotics Research*, 31(3):360–375, 2011.

G. Konidaris, L. Kaelbling, and T. Lozano-Perez. Constructing symbolic representations for high-level planning. In *Proceedings of the Twenty-Eighth Conference on Artificial Intelligence (AAAI)*, 2014.

S. Krishnan, A. Garg, R. Liaw, L. Miller, F. T. Pokorny, and K. Goldberg. HIRL: hierarchical inverse reinforcement learning for long-horizon tasks with delayed rewards. *CoRR*, abs/1604.06508, 2016. URL `http://arxiv.org/abs/1604.06508`.

O. Kroemer, H. van Hoof, G. Neumann, and J. Peters. Learning to predict phases of manipulation tasks as hidden states. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 4009–4014, 2014.

O. Kroemer, C. Daniel, G. Neumann, H. van Hoof, and J. Peters. Towards learning hierarchical skills for multi-phase manipulation tasks. In *Proceedings of International Conference on Robotics and Automation (ICRA)*, pages 1503 – 1510, 2015.

K. Kronander, M. Khansari, and A. Billard. Incremental motion learning with locally modulated dynamical systems. *Robotics and Autonomous Systems*, 70(C):52–62, 2015.

V. Kuleshov and O. Schrijvers. Inverse game theory: Learning utilities in succinct games. In *Proceedings of the International Conference on Web and Internet Economics*, 2015.

D. Kulić, W. Takano, and Yoshihiko Nakamura. Incremental learning, clustering and hierarchy formation of whole body motion patterns using adaptive hidden markov chains. *The International Journal of Robotics Research*, 27: 761–784, 2008.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Y. Kuniyoshi, M. Inaba, and H. Inoue. Learning by watching: extracting reusable task knowledge from visual observation of human performance. *IEEE Transactions on Robotics and Automation*, 10(6):799–822, 1994.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2001.

F. Lagriffoul, D. Dimitrov, J. Bidot, A. Saffiotti, and L. Karlsson. Efficiently combining task and motion planning using geometric constraints. *The International Journal of Robotics Research*, 33(14):1726–1747, 2014.

M. Laskey, J. Lee, W. Hsieh, R. Liaw, J. Mahler, R. Fox, and K. Goldberg. Iterative noise injection for scalable imitation learning. *arXiv preprint*, 2017.

Y. LeCun, U. Muller, J. Ben, E. Cosatto, and B. Flepp. Off-road obstacle avoidance through end-to-end learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

A. Lee, H. Lu, A. Gupta, S. Levine, and P. Abbeel. Learning force-based manipulation of deformable objects from multiple demonstrations. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015a.

A. X. Lee, M. A. Goldstein, S. T. Barratt, and P. Abbeel. A non-rigid point and normal registration algorithm with applications to learning from demonstrations. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015b.

D. Lee and Y. Nakamura. Mimesis model from partial observations for a humanoid robot. *The International Journal of Robotics Research*, 2009.

D. Lee and C. Ott. Incremental kinesthetic teaching of motion primitives using the motion refinement tube. *Autonomous Robots*, 2011.

D. Lee, C. Ott, and Y. Nakamura. Mimetic communication model with compliant physical contact in human-humanoid interaction. *The International Journal of Robotics Research*, 29:1684–1704, 2010.

A. Lemme, Y. Meirovitch, M. Khansari-Zadeh, T. Flash, A. Billard, and J. J. Steil. Open-source benchmarking for learned reaching motion generation in robotics. *Paladyn, Journal of Behavioral Robotics*, 6(1), 2015.

T. Lens, J. Kunz, O. v. Stryk, C. Trommer, and A. Karguth. Biorob-arm: A quickly deployable and intrinsically safe, light- weight robot arm for service robotics applications. In *International Symposium on Robotics (ISR)*, pages 1–6, 2010.

S. Levine and P. Abbeel. Learning neural network policies with guided policy search under unknown dynamics. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

S. Levine and V. Koltun. Continuous inverse optimal control with locally optimal examples. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 41–48, 2012.

S. Levine, Z. Popovic, and V. Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 19–27, 2011.

S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1): 1334–1373, 2016.

R. Lioutikov, G. Neumann, G. Maeda, and J. Peters. Learning movement primitive libraries through probabilistic segmentation. *The International Journal of Robotics Research (IJRR)*, 36(8):879–894, 2017.

Y. Liu, A. Gupta, P. Abbeel, and S. Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. *arXiv*, 2017.

M. Lopes, F. Melo, and L. Montesano. *Active Learning for Reward Estimation in Inverse Reinforcement Learning*, pages 31–46. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-04174-7. .

T. Lozano-Perez, J. L. Jones, E. Mazer, and P. A. O'Donnell. Task-level planning of pick-and-place robot motions. *Computer*, 22(3):21–29, 1989.

L. Lukic, J. Santos-Victor, and A. Billard. Learning robotic eye-arm-hand coordination from human demonstration: a coupled dynamical systems approach. *Biological Cybernetics*, 108(2):223–248, 2014.

G. Maeda, G. Neumann, M. Ewerton, L. Lioutikov, O. Kroemer, and J. Peters. Probabilistic movement primitives for coordination of multiple human-robot collaborative tasks. *Autonomous Robots*, 2016.

G. Maeda, M. Ewerton, T. Osa, B. Busch, and J. Peters. Active incremental learning of robot movement primitives. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2017.

P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, 1936.

S. Manschitz, J. Kober, M. Gienger, and J. Peters. Learning movement primitive attractor goals and sequential skills from kinesthetic demonstrations. *Robotics and Autonomous Systems*, 74:97–107, 2015.

J. Maryniak, E. Ładyżyńska-Kozdraś, and S. Tomczak. Configurations of the Graf-Boklev (V-style) ski jumper model and aerodynamic parameters in a wind tunnel. *Human Movement*, 10(2):130–136, 2009.

H. Miyamoto, S. Schaal, F. Gandolfoc, H. Gomi, Y. Koike, R. Osu, E. Nakano, Y. Wada, and M. Kawato. A kendama learning robot based on bidirectional theory. *Neural Networks*, 9(8):1281–1302, 1996.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

P. Moylan and B. Anderson. Nonlinear regulator theory and an inverse optimal control problem. *IEEE Transactions on Automatic Control*, 18(5): 460–465, 1973.

K. Mülling, O. Kroemer J. Kober and, and J. Peters. Learning to select and generalize striking movements in robot table tennis. *The International Journal of Robotics Research*, 32:263–279, 2013.

A. Nair, D. Chen, P. Agrawal, P. Isola, P. Abbeel, J. Malik, and S. Levine. Combining self-supervised learning and imitation for vision-based rope manipulation. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2017.

S. Nakaoka, A. Nakazawa, F. Kanehiro, K. Kaneko, M. Morisawa, H. Hirukawa, and K. Ikeuchi. Learning from observation paradigm: leg task models for enabling a biped humanoid robot to imitate human dances. *The International Journal of Robotics Research*, 26(8):829–844, 2007.

S. Natarajan, G. Kunapuli, K. Judah, P. Tadepalli, K. Kersting, and J. Shavlik. Multi-agent inverse reinforcement learning. In *Proceedings of the ninth International Conference on Machine Learning and Applications (ICMLA)*, pages 395–400. IEEE, 2010.

G. Neu and C. Szepesvári. Training parsers by inverse reinforcement learning. *Machine learning*, 77(2-3):303–337, 2009.

A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 663–670, 2000.

D. Nguyen-Tuong and J. Peters. Model learning for robot control: a survey. *Cognitive Processing*, 12(4):319–340, 2011.

S. Niekum, S. Osentoski, G. Konidaris, S. Chitta, B. Marthi, and A. G. Barto. Learning grounded finite-state representations from unstructured demonstrations. *The International Journal of Robotics Research*, 34:131–157, 2014.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 271–279, 2016.

J. Oh, X. Guo, H. Lee, R. Lewis, and S. Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2845–2853, 2015.

T. Okamoto, T. Shiratori, S. Kudoh, S. Nakaoka, and K. Ikeuchi. Toward a dancing robot with listening capability: keypose-based integration of lower-, middle-, and upper-body motions for varying music tempos. *IEEE Transactions on Robotics*, 30(3):771–778, 2014.

T. Osa, N. Sugita, and M. Mitsuishi. Online trajectory planning in dynamic environments for surgical task automation. In *Proceedings of Robotics: Science and Systems (R:SS)*, 2014.

T. Osa, A. M. Ghalamzan E., R. Stolkin, R. Lioutikov, J. Peters, and G. Neumann. Guiding trajectory optimization by demonstrated distributions. *IEEE Robotics and Automation Letters (RA-L)*, 2(2):819–826, 2017a.

T. Osa, N. Sugita, and M. Mitsuishi. Online trajectory planning and force control for automation of surgical tasks. *IEEE Transactions on Automation Science and Engineering*, 2017b.

A. Paraschos, C. Daniel, J. Peters, and G. Neumann. Probabilistic movement primitives. In *Proceedings of Advances in Neural Information Processing Systems 26*, 2013.

R. Parent. *Computer animation: algorithms and techniques.* Morgan Kaufmann, 2002.

S. Y. Park and A. K. Bera. Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics*, 150(2):219–230, 2009.

P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal. Learning and generalization of motor skills by learning from demonstration. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

D. A. Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems (NIPS)*, 1988.

P. Poupart and N. Vlassis. Model-based Bayesian reinforcement learning in partially observable domains. In *Proceedings of the Tenth International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2008.

L. R. Rabiner. A tutorial on hidden markov models and selected appli- cations in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

M. Racca, J. Pajarinen, A. Montebelli, and V. Kyrki. Learning in-contact control strategies from demonstration. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 688–695. IEEE, 2016.

R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine. Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration. *arXiv*, 2017.

D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *Proceedings of the International Joint Conference on Artifical Intelligence (IJCAI)*, pages 2586–2591, 2007.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.

N. Ratliff, D. Bradley, J. A. Bagnell, and J. Chestnutt. Boosting structured prediction for imitation learning. In *Advances in Neural Information Processing Systems 19*, 2006a.

N. Ratliff, D. Silver, and J. A. Bagnell. Learning to search: Functional gradient techniques for imitation learning. *Autonomous Robots*, 27:25–53, 2009.

N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *Proceedings of the international conference on Machine learning (ICML)*, pages 729–736, 2006b.

S. Ross and J. A. Bagnell. Efficient reductions for imitation learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

S. Ross and J. A. Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *Arxiv preprint*, 2014.

S. Ross, G. J. Gordon, and J. A. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

S. Ross, N. Melik-Barkhudarov, K. S. Shankar, A. Wendel, D. Dey, J. A. Bagnell, and M. Hebert. Learning monocular reactive uav control in cluttered natural environments. In *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pages 1765–1772, May 2013. .

L. Rozo, Silvério J., S. Calinon, and D. Caldwell. Learning controllers for reactive and proactive behaviors in human-robot collaboration. *Frontiers in Robotics and AI*, pages 1–11, 2016.

S. Russell. Learning agents for uncertain environments (extended abstract). In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998.

J. Rust. *Handbook of Econometrics*, chapter Structural estimation of Markov decision processes, pages 3082–3139. Elsevier, 1994.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

C. Sammut, S. Hurst, D. Kedzier, and D. Michie. Learning to fly. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 385–393, 1992.

S. Schaal. Learning from demonstration. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1040–1046, 1997.

S. Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3(6):233 – 242, 1999.

S. Schaal and C. Atkeson. Constructive incremental learning from only local information. *Neural Computation*, 10(8):2047–2084, 1998.

S. Schaal, J. Peters, J. Nakanishi, and A. Ijspeert. Learning movement primitives. In *Proceedings of the International Symposium on Robotics Research (ISRR)*, 2004.

J. G. Schneider. Exploiting model uncertainty estimates for safe dynamic control learning. In *Advances in Neural Information Processing Systems (NIPS)*, 1997.

J. Schulman, J. Ho, C. Lee, and P. Abbeel. Learning from demonstrations through the use of non-rigid registration. In *Proceedings of the International Symposium on Robotics Research (ISRR)*, 2013.

J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

R. Serfozo. *Basics of Applied Stochastic Processes*. Springer Science & Business Media, 2009.

P. Sermanet, K. Xu, and S. Levine. Unsupervised perceptual rewards for imitation learning. In *Proceedings of Robotics and Science and Systems (R:SS)*, 2017.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

K. Shiarlis, J. Messias, and S. Whiteson. Inverse reinforcement learning from failure. In *Proceedings of the International Conference on Autonomous Agents & Multiagent Systems*, pages 1060–1068. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

A. Shukla and A. Billard. Coupled dynamical system based arm-hand grasping model for learning fast adaptation strategies. *Robotics and Autonomous Systems*, 60(3):424–440, 2012.

D. Silver, J. A. Bagnell, and A. Stentz. Learning from demonstration for autonomous navigation in complex unstructured terrain. *The International Journal of Robotics Research*, 29(12):1565–1592, 2010.

D. Silver, J. A. Bagnell, and A. Stentz. Active learning from demonstration for robust autonomous navigation. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 200–207, 2012.

D. Silver, J. A. Bagnell, and A. Stentz. Learning autonomous driving styles and maneuvers from expert demonstration. In *Experimental Robotics: The 13th International Symposium on Experimental Robotics (ISER)*, pages 371–386, 2013.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587): 484–489, 2016.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

B. Stadie, P. Abbeel, and I. Sutskever. Third person imitation learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

M. Sugiyama. *Introduction to Statistical Machine Learning*. Morgan Kaufmann, 2015.

M. Sugiyama, M. Kawanabe, and P. L. Chui. Dimensionality reduction for density ratio estimation in high-dimensional spaces. *Neural Networks*, 23 (1):44–59, 2010.

M. Sugiyama, H. Hachiya, and T. Morimura. *Statistical Reinforcement Learning: Modern Machine Learning Approaches*. Chapman & Hall/CRC, 2013.

W. Sun, A. Venkatraman, G. Gordon, B. Boots, and J. A. Bagnell. Deeply aggrevated: Differentiable imitation learning for sequential prediction. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

R. Sutton and A. Barto. *Reinforcement learning: An introduction*. The MIT Press, 1998.

R. S. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1-2):181–211, 1999.

C. Szepesvari. Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, 2010. .

W. Takano and Y. Nakamura. Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. *The International Journal of Robotics Research*, 34:1314–1328, 2015.

W. Takano and Y. Nakamura. Real-time unsupervised segmentation of human whole-body motion and its application to humanoid robot acquisition of motion symbols. *Robotics and Autonomous Systems*, 75:260–272, 2016.

W. Takano and Y. Nakamura. Planning of goal-oriented motion from stochastic motion primitives and optimal controlling of joint torques in whole-body. *Robotics and Autonomous Systems*, 91:226–233, 2017.

V. Tangkaratt, N. Xie, and M. Sugiyama. Conditional density estimation with dimensionality reduction via squared-loss conditional entropy minimization. *Neural Computation*, 27(1):228–254, 2015.

B. Taskar. *Learning structured prediction models: a large margin approach*. PhD thesis, Stanford University, 2005.

Y. Tassa, T. Erez, and E. Todorov. Synthesis and stabilization of complex behaviors through online trajectory optimization. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4906–4913, 2012.

G. Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58 – 68, 1995.

E. Todorov and W. Li. A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the American Control Conference*, 2005.

I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

A. Ude, C. G. Atkeson, and M. Riley. Programming full-body movements for humanoid robots by observation. *Robotics and Autonomous Systems*, pages 93–108, 2004.

J. van den Berg, S. Miller, D. Duckworth, H. Hu, A. Wan, X. Y. Fu, K. Goldberg, and P. Abbeel. Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2074–2081, 2010.

V. N. Vapnik. *Statistical learning theory.* John Wiley & Sons, 1998.

A. Venkatraman, M. Hebert, and J. A. Bagnell. Improving multi-step prediction of learned time series models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 3024–3030, 2015.

A. Venkatraman, R. Capobianco, L. Pinto, M. Hebert, D. Nardi, and J. A. Bagnell. Improved learning of dynamics models for control. In *Proceedings of the International Symposium on Experimental Robotics (ISER)*, 2016.

S. Vijayakumar and S. Schaal. Locally weighted projection regression: An o(n) algorithm for incremental real time learning in high dimensional space. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1079–1086, 2000.

S. Vijayakumar, A. D'Souza, and S. Schaal. Incremental online learning in high dimensions. *Neural Computation*, 17:2602–2634, 2005.

K. Waugh, B. D. Ziebart, and J. A. Bagnell. Computational rationalization: The inverse equilibrium problem. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1169–1176, 2011.

T. H. Wen, M. Gašić, N. Mrkšić, P. H. Su, D. Vandyke, and S. Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1711–1721. Association for Computational Linguistics, 2015.

B. Widrow and F. W. Smith. Pattern recognising control systems. *Computer and Information Sciences Clever Hume Press*, 1964.

M. Wiering and M. van Otterlo, editors. *Reinforcement Learning: State-of-the-Art.* Springer, 2012.

S. Z. Yu. Hidden semi-markov models. *Artificial Intelligence*, 174(2):215–243, 2010.

B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the IEEE International Conference on Data Mining*, pages 435–442, 2003.

B. D. Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. PhD thesis, University of Washington, 2010.

B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI)*, pages 1433–1438, 2008.

B. D. Ziebart, J. A. Bagnell, and A. K. Dey. The principle of maximum causal entropy for estimating interacting processes. *IEEE Transactions on Information Theory*, 59(4):1966–1980, 2013.

M. Zucker, N. Ratliff, M. Stolle, J. Chestnutt, J Andrew Bagnell, C. G. Atkeson, and J. Kuffner. Optimization and learning for rough terrain legged locomotion. *The International Journal of Robotics Research*, 30(2):175–191, 2011.

M. Zucker, N. Ratliff, A. D. Dragan, M. Pivtoraiko, M. Klingensmith, C. M. Dellin, J. A. Bagnell, and S. S. Srinivasa. Chomp: covariant hamiltonian optimization for motion planning. *The International Journal of Robotics Research*, 32:1164–1193, 2013.