Annual Reviews in Control xxx (xxxx) xxx

Contents lists available at ScienceDirect

Annual Reviews in Control

journal homepage: www.elsevier.com/locate/arcontrol

From inverse optimal control to inverse reinforcement learning: A historical review

Nematollah Ab Azar*, Aref Shahmansoorian, Mohsen Davoudi

Department of Electrical Engineering, Imam Khomeini International University (IKIU), Qazvin, Iran

A R T I C L E I N F O

Article history:
Received 11 October 2019
Revised 21 May 2020
Accepted 16 June 2020
Available online xxx

Keywords:
Inverse optimal control
Inverse reinforcement learning
Learning from demonstration
Imitation learning

A B S T R A C T

Inverse optimal control (IOC) is a powerful theory that addresses the inverse problems in control systems, robotics, Machine Learning (ML) and optimization taking into account the optimal manners. This paper reviews the history of the IOC and Inverse Reinforcement Learning (IRL) approaches and describes the connections and differences between them to cover the research gap in the existing literature. The general formulation of IOC/IRL is described and the related methods are categorized based on a hierarchical approach. For this purpose, IOC methods are categorized under two classes, namely classic and modern approaches. The classic IOC is typically formulated for control systems, while IRL, as a modern approach to IOC, is considered for machine learning problems. Despite the presence of a handful of IOC/IRL methods, a comprehensive categorization of these methods is lacking. In addition to the IOC/IRL problems, this paper elaborates, where necessary, on other relevant concepts such as Learning from Demonstration (LfD), Imitation Learning (IL), and Behavioral Cloning. Some of the challenges encountered in the IOC/IRL problems are further discussed in this work, including ill-posedness, non-convexity, data availability, nonlinearity, the curses of complexity and dimensionality, feature selection, and generalizability.

Contents

1. Introduction

Inverse Optimal Control (IOC) (Kalman, 1964) and Inverse Reinforcement Learning (IRL) (Ng & Russell, 2000) are two well-known inverse-problem frameworks in the fields of control and machine learning. Although these two methods follow similar goals, they differ in structure. The IOC aims to reconstruct an objective function given the state/action samples assuming a stable control system, while the IRL recovers an objective function using expert demonstration assuming that the expert behavior is optimal. “Imitation Learning (IL)” or what is known as “Learning from Demonstration (LfD)” (Schaal, 1997) is a type of learning problem in which an agent tries to learn a task using the given demonstrated data including the path trajectory or state-control measurements. Two approaches have been proposed to the IL problem, including Behavioral Cloning (BC) (Bain & Sommut, 1999)

---

* Corresponding author.
E-mail addresses: n.abazar@edu.ikiu.ac.ir (N. Ab Azar), shahmansoorian@eng.ikiu.ac.ir (A. Shahmansoorian), davoudi@eng.ikiu.ac.ir (M. Davoudi).

https://doi.org/10.1016/j.arcontrol.2020.06.001
1367-5788/© 2020 Elsevier Ltd. All rights reserved.

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

and the IOC/IRL. In this paper, IOC is investigated under two categories, namely classic and modern IOC. The classic approach to IOC is typically formulated for control systems based on the stabilization control laws, passivity condition, reconstruction of Control Lyapunov Function (CLF), Lagrange Multipliers and Proportional-Integral-Derivative (PID) gains, while IRL, introducing the modern approach to IOC, is considered for machine learning problems. Several approaches have been proposed for solving the IOC problem, including deterministic, stochastic, and Bayesian methods, Markov Decision Process (MDP), and related functions including regression, clustering, matching, regularization, and optimization functions as well as Dynamic Programming (DP), Linear Programming (LP), Quadratic Programming (QP), Semi-Definite Programming (SDP), Convex Optimization (CO), and Polynomial Optimization (PO) techniques. Being multidisciplinary and having contributions from one another, the mentioned topics are difficult to distinguish, highlighting the importance of categorization of them to better understand and study the IOC/IRL problems.

Acronyms

Adaptive Critic Designs (ACDs)
Apprenticeship Learning (AL)
Approximate Dynamic Programming (ADP)
Approximate Linear Programming (ALP)
Behavioral Cloning (BC)
Bayesian IRL (BIRL)
Bayesian nonparametric IRL(BNIRL)
Compatible Reward IRL (CRIRL)
Continuous Inverse Optimal Control (CIOC)
Control Lyapunov Function (CLF)
Convex Optimization (CO)
Deep Apprenticeship Q-Network (DAQN)
Deep Apprenticeship Reward Network (DARN)
Deep Gaussian Process IRL (DGP-IRL)
Deep Maximum Entropy IRL(DMaxEntIRL)
Dynamic Programming (DP)
Extended Kalman Filter (EKF)
Feature Expectation Matching (FEM)
Feature construction IRL (FIRL)
Generative Adversarial Imitation Learning (GAIL)
Generative Adversarial Maximum Entropy (GAMaxEntIRL)
Generative Adversarial Network (OptionGAN)
Gaussian Process IRL (GPIRL)
Imitation Learning (IL)
Input-to-State Stability (ISS)
Inverse Reinforcement Learning (IRL)
Inverse Optimal Control (IOC)
IOC Differential Dynamic Programming (IOCDDP)
Karush-Kohn-Tucker (KKT)
Learning from Demonstration (LfD)
Learning to Search (LEARCH)
Linear Matrix Inequality (LMI)
Linear-Quadratic (LQ)
Linear-Quadratic Regulator (LQR)
Linearly-solvable MDPs (LMDPs)
Machine Learning (ML)
Markov Decision Process (MDP)
Maximum-Entropy IRL (MaxEntIRL)
Maximum Entropy Semi-Supervised IRL (MESS IRL)
Maximum Likelihood IRL (MLIRL)
Max-Margin Planning (MMP)
Multiplicative Weight Apprenticeship Learning algorithm (MWAL)
Partially Observable Markov Decision Process (POMDP) IRL
Path Integrals IRL (PI-IRL)
Policy Gradient (PG)
Policy Gradient IRL (PGIRL)
Policy Matching (PM)
Polynomial Optimization (PO)
Proportional-Integral-Derivative (PID)
Proximal Policy Optimization (PPO)
Quadratic Programming (QP)
Reinforcement Learning (RL)
Relative entropy IRL (REIRL)
Semi-Definite Programming (SDP)
Structured Classification IRL (SCIRL)
Trust Region Policy Optimization (TRPO)

This review is significant because of the followings:

1. Providing a historical view to IOC and IRL and recognizing the connections and differences between them;
2. Arranging the two methods under a hierarchy to categorize related methods; this is especially important as it is yet to be properly described based on a comprehensive hierarchal structure; and
3. Discussing the challenging encountered in the IRL/IOC problems.

The categorization of these methods has been based on understanding their differences and similarities which might be arisen from the following reasons:

1. System and environment model: A model of the system and environment shows how the current state transfers to the next state. It may be linear or non-linear, dynamic or static, deterministic or stochastic, discrete or continuous, globally or locally stable, discrete or continuous, simple or complex, ordered or random, analytic or numerical, constructive or nonconstructive, stable or unstable, and model-based or model-free. In addition to the above properties, the model of environment may be deterministic or stochastic, fully or partially observable, gridworld or real-world, and convex or non-convex. Therefore, any assumption on the system or environment model requires different approaches to solve the LfD problem.
2. Demonstration: There are various sources for demonstration, including the teacher, learner, and/or observer demonstration. Based on the availability of related measurements, one may prefer to choose the most appropriate method for data collection. For example, for learning an acrobat, the demonstration data can be either captured from sensors attached to the joints of an expert demonstrator (teacher), recorded by a camera as an external source, or provided by the sensors attached to the learner himself/herself. Also, a model of the demonstrated trajectories may be deterministic (i.e. the state/action trajectories) or stochastic (i.e. the probabilistic densities of the given trajectories). Being either partial or complete, the observations can be in the form of states or control inputs that are attained by teacher/learner/expert.
3. Control law or policy: A control law is a function that determines a control value for a state. Similarly, a policy maps a state to an action. A control law or policy may be either deterministic or stochastic, stationary or non-stationary, linear or non-linear, feed-back or open-loop (i.e. feedback-free), and hierarchical or monolithic. A policy or control law can be obtained in several ways based upon different assumptions.
4. Construction of cost or reward function: Based on the shape of the objective or reward function, different problem-solving methods may be used, including linear, quadratic, cubic, Gaussian, feature-based, convex and non-convex methods. To construct an appropriate cost function, one must select the smallest subset of the features that can represent the model or value function. A feature is an individual measurable property or characteristic of the phenomenon being observed (Bishop, 2006). Accordingly, the properties of an object (e.g. color, size, and shape) are the features of the object. The constraints are limitations imposed on the state or control variables or existing obstacles in the environment. For example, a robot must remain within predefined limits on the velocity, angle, and torques and respect its kinematic constraints. An optimal control problem is formed based upon an objective function subjected to the control system dynamics, and a set of constraints or features.
5. The system behavior: The system behavior is key to select the appropriate approach to LfD. Accordingly, an LfD problem where the system is stable can be solved through the classic approach to IOC. In this case, stability is an internal behavior of the system, making the problem solvable

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

through a classic approach. Assuming that the system behaves optimally implies that the system performs behaves like an expert to have the tasks performed optimally (e.g. optimal policy). The classic approach can not be useful in this case, because “the optimality does not imply stability” (Kalman, 1964). Even though some optimal systems are stable, but there is commonly not enough information on the system dynamics. The IRL provides robust approaches to solving such problems.

Based on the above-mentioned items, the choice of the appropriate approach to an LfD problem depends on the types of demonstration, objective function, selected features, and system and environmental model, so that such a problem cannot be efficiently solved if appropriate assumptions are not considered.

This paper reviews various IOC/IRL methods, problems, and challenges. Chapters 2 describes the history of the IOC/IRL. The general formulations of IOC and IRL are explained in Chapter 3. The categorization of the IOC approaches is addressed in Chapter 4. Chapters 5 presents several applicable and useful methods in the fields of IOC and IRL. Finally, Chapter 6 presents a discussion on the existing challenges regarding the IOC/IRL methods.

2. History of IOC/IRL methods

Inverse optimal control is a powerful theory that is used to address the inverse problem in control systems, robotics, machine learning, and optimization taking into account the so-called optimal manners. Abel was the first to mathematically study an inverse mechanical problem for finding the curve of an unknown path in 1826 (Yaman, Yakhno & Potthast, 2013). Being closely associated with optimal control, the IOC has gone through a long history of science and technology. Optimal control has been a subject of interest for researchers for over 300 years (Sussmann & Willems, 1997). Originally, optimal control has been a significant and widely applied part of the calculus of variations since the 17th century before it was developed further by Euler and Lagrange in the 18th century and later in the 19th century by Jacobi, Hamilton, and Weierstrass. During the 20th century, remarkable developments were achieved in this branch of mathematical science; these were contributed by Bolza (1909), McShane (1939), Bliss (1946), Bellman (1957) and Pontryagin et al. (1961), among others (Bryson, 1996). The problem of optimizing linear-stationary control systems subject to quadratic performance indices and bounded control effort constraints was first proposed and studied by Letov (1960). Following the Letov problem, Rekasius and Hsia (1964) developed the proposed solution by implementing necessary and sufficient conditions for the existence of the optimal control laws of the saturation type. Studying the inverse problems by emphasizing the forms to determine Lagrangians given a family of curves has been a sub-chapter of the calculus of variations (Bolza, 1909; Akhiezer, 1962). Bellman and Kalaba (1963) were the first to introduce the inverse problem in the context of dynamic programming and automatic control to find the criterion function given an optimal policy and the descriptive equations.

As the first formulation of the IOC, Kalman (1964) devised a method to formulate, study, and resolve the inverse problem of optimal control theory in an attempt to find all performance indices given a control law under the assumptions of linear constant plant and control law, measurable state variables, and quadratic loss functions with constant coefficients and a single control variable. Obermayer and Muckler (1965) developed a special method for computing optimal performance weighting coefficients for manual control. Anderson (1966) generalized the Kalman problem to a multi-input, time-invariant case (Jameson & Kreindler, 1973). Thau (1967) proposed the IOC theory corresponding to the Lure problem for a class of non-linear control systems to determine the performance criteria for which the given control law was optimum. Kurz (1969) presented a solution to address the IOC problem for deterministic growth paths for economic applications. Park and Lee (1975) proposed a performance index for economic stabilization policy using the inverse optimal control given a closed-loop system and a known control policy. Willems and Van De Voorde (1977) used the inverse control problem for a linear discrete-time system and discussed its differences with a continuous-time system. Wei and Shieh (1979) presented a new method to use IOC for a class of multivariable control systems to determine the block-weighting matrices of the quadratic performance index from given control specifications.

In connection with the DP and parallel to the application of inverse problems in the calculus of variations, Bellman (1970) highlighted a number of the advantages of the DP for classic IOC based on stochastic and adaptive processes. Iwamoto (1976) developed a new inverse theory of DP and applied it to mathematical programming problems and discrete-time control and allocation processes. Casti (1980) elaborated on the IOC problem in the framework of DP while respecting the necessary and sufficient conditions for linear quadratic problems. Chang (1988) proposed a stochastic IOC problem in the framework of DP. Bellman and Dreyfus (1959) proposed to use functional approximations as a way to overcome the challenge of approximating value functions. This method was later continued by Schweitzer and Seidmann (1985) in the modern era for value function approximations. Werbos (1977) introduced an approach for Approximate Dynamic Programming(ADP) that was later called Adaptive Critic Designs (ACDs). In its literature, there are several synonyms or branches for ACDs including Asymptotic Dynamic Programming (Saeks, Cox, Mathia & Maren, 1997), Adaptive Dynamic Programming (Murray, Cox, Lendaris & Saeks, 2002), Cost approximate dynamic programming (De Farias & Van Roy, 2006), Heuristic Dynamic Programming (Werbos, 1992; Lendaris & Paintz, 1997), Neuro-Dynamic Programming (Bertsekas and Tsitsiklis, 1995; Bertsekas and Tsitsiklis, 1996; Powell, 2008), Neural Dynamic Programming (Si et al., 2004), and Reinforcement Learning (Sutton & Barto, 1998; Lewis & Liu, 2013). ADP has paid much attention from many researchers to obtain approximate solutions of the HJB equation (Al-Tamimi, Lewis & Abu-Khalaf, 2008; Wang, Zhang and Liu, 2009) because it is a powerful technique to solve large scale discrete-time multistage stochastic control processes, i.e., complex Markov Decision Processes (MDPs) (Mes & Rivera, 2017) and also it is a way to avoid the curse of dimensionality by using the methodology of LP-based ADP called Approximate Linear Programming (ALP). Reinforcement Learning (RL), as a method that initially seems is closely related to AD, has played an important and undeniable role in the development of machine learning. Inverse reinforcement learning is one of the achievements of this branch of ML, has been developed based on practical needs to solve inverse problems in learning topics. Therefore, to understanding IRL, studying RL, the inverse problem in control theory, and also DP is essential and inevitable.

Concerning the linear-quadratic (LQ) problems, the quadratic weights are usually adjusted by a trial-and-error mechanism to reach the desired control specifications. Proposed by Kalman (1964), the inverse LQ problem was a single-input infinite-time problem. The problem was later on approached in the time domain by Jameson and Kreindler (1973). Molinari (1973) provided a detailed survey and extension of certain properties of the stable regulator problem providing the relationship between the time-domain and frequency-domain solutions to solve the inverse problem of whether a given state feedback control law is optimal. A multi-input inverse LQ problem was proposed by Anderson & Moore (1989). Kawasaki and Shimemura (1983) proposed a new procedure to determine a weighting matrix for an LQ problem within the framework of the pole assignment problem in such a way that all poles of the closed-loop system were placed in the desired region for good response and stability. Fujii and Narazaki (1984) determined a new geometric condition, i.e. a complete set of optimality conditions, for the IOC problem. Fujii (1987) presented an approach to the LQ design in the framework of the inverse regulator problem. Mehdi, Darouach and Zasadzinski (1994) used the inverse optimal regulator problem to address a discrete-time LQ design. Sugimoto (1998) followed an inverse approach to the pole-placement problem using LQ regulators. More recently, the Linear Matrix Inequality (LMI) and optimization tools were used by Boyd, El Ghaoui, Feron and Balakrishnan (1994) and Priess et al. (2014) to calculate solutions for Linear-Quadratic Regulator (LQR) problems. Krstic and Tsiotras (1999) used IOC to solve the corresponding LQR problem of stabilizing a rigid spacecraft.

Regarding the non-linear IOC, Casti (1974) presented some equations for a specific feedback control law and presented a generalization form to extend a linear regulator case to non-quadratic criteria by appropriate parameterization of the cost function. Numerous non-linear approaches to the IOC have been proposed for using the passivity condition (Moylan & Anderson, 1973; Ortega et al., 1990), receding horizon feedback control (RHFC) (Chen & Shaw, 1982), robust design of IOC (Freeman & Kokotovic, 1996), constructive non-linear control (Sepulchre, Jankovic & Kokotovic, 1997), global attitude stabilization (Osipchuk, Bharadwaj & Mease, 1997), construction of an optimal feedback control law (Krstic & Tsiotras, 1999), and determination of control- and state-space trajectories (Radoslav, 1988).

Focusing on the ill-posedness problem, it should be noted that an inverse problem is typically more difficult to solve numerically than a forward problem, because of the ill-posedness of the former. As an ill-posed problem, the IOC is also facing such difficulty. Regularization has been robustly used to overcome the ill-posedness of the inverse problems. Yeh (1986) reviewed the methods of parameter identification for groundwater hydrology and addressed the ill-posedness problem related to the inverse problem. Busby and Trujillo (1997) approached the optimal regularization for an inverse dynamics problem. Vito, Rosasco, Caponnetto, Giovannini and Odone (2005) analyzed the connection between the theory of statistical learning and the theory of ill-posed problems. Levine, Popovic and Koltun (2011) used a regularization method for non-linear IRL. Pauwels, Henrion and Lasserre (2014), Pauwels, Henrion & Lasserre (2016) and Rouot and Lasserre (2017) used the regularization concept for a polynomial-optimization IOC problem.

In connection with the adaptive control, Widrow (1987) introduced an adaptive inverse identification process to obtain a stable controller by adjusting its parameters assuming that the system is affected by internal noise and the plant is a non-minimum phase. Krstic, Kanellakopoulos and Kokotovic (1995) addressed the non-linear adaptive control design. Ortega, Rodriguez and Espinosa (1990) studied the problem of adaptive stabilization of non-linear systems. Widrow and Plett (1996) used adaptive filtering-based inverse control to achieve feedforward control considering linear and non-linear, MIMO, and SISO plants. Fausz, Chellaboina and Haddad (2000) proposed a Lyapunov-based inverse optimal adaptive control-system for non-linear uncertain systems. Plett (2003) combined adaptive inverse control and neural network for controlling linear and non-linear systems.

Concurrently with the development of the IOC theory, practical applications and methodologies were presented for several real fields, including, but not limited to, the application of the IOC theory for aircraft stabilization system (Porter and Woodhead, 1970), inverse design and control of heat conduction (Dulikravich, 1988), robotics control (Spong & Ortega, 1990; Madhavan & Singh, 1991), inverse optimal control for a rigid spacecraft (Krstic & Tsiotras, 1997), control of piezoelectric actuators with hysteresis operators (Kuhnen & Janocha, 1999), multiple paired forward and inverse models for motor control (Wolpert & Kawato, 1998), constructive Lyapunov control design for turbo-charged diesel engines (Jankovic & Kolmanovsky, 2000), hysteresis and creep control (Krejci & Kuhnen, 2001), aircraft flight path reconstruction (Blajer, Goszczyński & Krawczyk, 2002), chaos control (Sanchez, Perez, Martinez & Chen, 2002; Zhang, Wang & Liu, 2004), and CLF construction for electric power conversion (Vega & Alzate, 2014).

The robust approach to the control was increasingly interested in the 90 s. Numerous research works have focused on the inverse problems under robust control theory, including a robust collocated control system for large flexible space structures (Arbel & Gupta, 1981), LQ differential games (Fujii & Khargonekar, 1988), robust inverse-optimal control for the Euler-Lagrange system (Park & Chung, 2000), inverse optimal robust stabilization problem for non-linear systems with disturbances in the context of CLF (Freeman & Kokotovic, 1996, 2008), and practicality of the robust IOC (Luo, Chu & Ling, 2005).

As an essential paradigm in control theory, the concept of Control Lyapunov Function (CLF) was introduced by Sontag (1983) and Artstein (1983) for designing stabilizing control in a nonconstructive form. To generalization of the Artstein's theorem, Sontag (1989) provided a constructive version of it explicitly exhibiting the stabilizing control feedback so-called Sontag formula. The idea of Control Lyapunov Function has had a profound effect on the stabilization theory in control systems. Extensive research works have been conducted on designing stabilizers and constructing CLFs for optimal control problems including but not limited to stabilization with bounded controls (Lin & Sontag, 1991), the connection between stability and optimality (Sepulchre et al., 1997), integral-input-to-state stabilization (Liberzon, Sontag & Wang, 1999), construction of optimal feedback control laws for optimal regulation using IOC (Krstic & Tsiotras, 1997), inverse optimal design of input-to-state stabilizing non-linear controllers (Krstic, 1998; Krstic & Li, 1998; Krstic & Tsiotras, 1999), inverse optimal H_∞ design (Maruyama & Fujita, 1999), IOC approach for chaos stabilization (Sanchez et al., 2002), and generalization of Sontag's formula (Curtis III, 2002; Shahmansoorian, 2009). Freeman and Primbs (1996) analyzed a control design method for non-linear systems based on CLFs and inverse optimality to recover the LQ optimal control. Deng and Krstić (1997) designed stabilizing control laws that were optimal regarding a significant cost function using stochastic CLF. Magni and Sepulchre (1997) derived an inverse optimal theory for several stability margins of the non-linear receding-horizon control schemes. Kogan (1997) focused on solving the inverse problems of min-max control and worst-case disturbance for linear discrete-time systems. Fausz et al. (2000) proposed a Lyapunov-based adaptive IOC system to explicitly characterize globally stabilizing disturbance rejection for non-linear uncertain systems considering the disturbances. Ornelas, Sanchez and Loukianov (2010) developed a discrete-time IOC scheme for trajectory tracking of non-linear systems. Li, Todorov and Liu (2011) developed a new algorithm to construct an estimated cost function to recover the original performance criterion for which this control law was optimal. Khansari-Zadeh and Billard (2014) extended the classical CLF-based control scheme in the context of learning to learn CLF from demonstrations by a combination of sampled states and corresponding feedbacks provided by the demonstrator. Similar work was published by Ravanbakhsh and Sankaranarayanan (2019) who focused on learning simple polynomial CLFs from counterexamples and demonstrations for non-linear dynamical systems. Rohrweck et al. (2015) presented an approach to approximate the solution of an infinite-horizon optimal control problem for a class of non-linear systems by constructing a CLF.

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

Almobaied, Eksin and Guzelkaya (2015), Almobaied, Eksin & Guzelkaya, 2018 constructed a CLF using the Extended Kalman Filter (EKF). Prasanna et al. (2019) applied the CLF-based IOC approach for non-linear inverse optimal control problems, and Huang, Ma and Vaidya (2019) designed a data-driven non-linear stabilization scheme using the Koopman operator.

Residual minimization is another approach to the IOC. Based on this approach, Keshavarz, Wang and Boyd (2011) proposed a method for approximating an objective function, given some demonstrated samples, by implementing the necessary and sufficient Karush-Kohn-Tucker (KKT) optimality conditions in a dual problem formulation. The residual approach was also utilized in subsequent works by Puydupin-Jamin et al. (2012), Aghasadeghi, Long and Bretl (2012), and Johnson, Aghasadeghi and Bretl (2013). Pauwels et al. (2014) used the Hamilton-Jacobi-Bellman sufficient optimality condition as a tool for analyzing an inverse problem and proposed a method for solving it using the polynomial optimization and linear matrix inequalities. In the same year, Claeys and Sepulchre (2014) implemented moment optimization techniques for reconstructing trajectories. Pauwels et al. (2016) and Rouot and Lasserre (2017) formulated an IOC problem in the framework of polynomial optimization via the Lasserre relaxation (Lasserre, 2001) by translating the original optimal control problem to a measure problem and, subsequently, relaxing the measure problem to a moment problem and finally transferring the moment problem to an LMI problem that could be readily solved by appropriate tools.

Sometimes called Learning from Demonstration (LfD) or programming by example, the Imitation Learning (IL) is a type of learning where an agent tries to learn a task based on some demonstrated data including the path trajectory or state-control measurements. Schaal (1997) elaborated on the LfD, which was previously referred to as programming by demonstration (Delson & West, 1996), seeing to doing (Bakker & Kuniyoshi, 1996), imitation learning (Hayes and Demiris, 1994), and teaching by showing (Kawato, Gandolfo, Gomi, & Wada, 1994; Montgomery, 1999) –research that has drawn significant deals of attention in the robotics and machine learning.

An IL method can be either model-based or model-free. Model-free methods learn a policy with no knowledge of the system dynamics. The main advantage of such methods is that the system dynamics are encoded only implicitly in the policies, while, as a major disadvantage, the prediction of future states is difficult by these techniques. A model-based IL method learns a policy that explicitly satisfies the system dynamics by leveraging its dynamics. However, learning/estimating system dynamics can be challenging when it comes to the model-based IL methods (Osa et al., 2018). Generally, model-based methods outperform the model-free techniques greatly. First, those are more sample-efficient, *i.e.* do not require as many environment interactions as needed by a model-free IL method. Second, the learned models can be transferred across tasks (Torabi, Warnell, & Stone, 2018).

Taking into account the type of demonstration, the research on the IL continued to cover new aspects of learning such as few-shot IL (Ravi & Larochelle, 2016), one-shot IL (Duan et al., 2017; Finn, Yu, Zhang, Abbeel, & Levine, 2017; Yu et al., 2018), zero-shot visual IL (Pathak et al., 2018), suboptimal demonstration (Zheng, Liu, & Ni, 2014; Choi, Lee, & Oh, 2019; Brown et al., 2019), and failed demonstration (Shiarlis, Messias, & Whiteson, 2016).

Based upon the type of the learning method, generally, this IL/LfD paradigm covers all related problems following either of two approaches: Behavioral Cloning (BC) and Inverse Reinforcement Learning (IRL). In the BC, demonstrated trajectories are used to learn a policy from the observed states, followed by applying the policy by the robot. In the IRL, the demonstrations are used to learn the teacher's reward function. Subsequently, a policy is learned that optimizes the learned cost function (Englert, Paraschos, Deisenroth, & Peters, 2013).

The BC is a direct IL paradigm in robotics and machine learning. Despite the numerous research works on the application of direct IL paradigm for the model-free problems (Bain & Sommut, 1999; Urbancic & Bratko, 1993; Byrne & Russon, 1998; Schaal, 1999; Billard and Matarić, 2000; Billard and Matarić, 2001; Schaal, Ijspeert, & Billard, 2003; Ijspeert et al., 2003; Calinon and Billard, 2007; Ross & Bagnell, 2010; Khansari-Zadeh & Billard, 2011; Ross, Gordon, & Bagnell, 2011; Ijspeert, Nakanishi, Hoffmann, Pastor, & Schaal, 2013; Paraschos, Daniel, Peters, & Neumann, 2013; Osa, Sugita & Mitsuishi, 2014; Takano & Nakamura, 2015; Maeda et al., 2017; Deniša et al., 2015; Ho & Ermon, 2016; Daftry, Bagnell, & Hebert, 2016; Torabi et al., 2018), only a few studies have specifically focused on the use of BC for the model-free problems (Ude, Atkeson, & Riley, 2004; Van Den Berg et al., 2010; Englert et al., 2013). This method is a supervised learning method in which the agent receives both the states and actions of a demonstrator as the training data and then uses a classifier or regressor to replicate the expert's policy (Ross & Bagnell, 2010).

As a secondary approach to the IL/LfD paradigm, IRL has been characterized by Russell (1998) to determine the reward function being optimized given 1. measurements of an agent's behavior over time, in a variety of circumstances, 2. measurements of the sensory inputs to that agent, and 3. a model of the physical environment (including the agent's body). Relying on this definition, Ng and Russell (2000) proposed the first IRL method to find the set of all reward functions for which a given policy was optimal or could explain the optimal behavior. Unfortunately, this formulation of the IRL problem was ill-posed due to the infinitely many reward functions that could be matched with the expert's demonstrated behavior. To address this issue, an alternative solution was to consider the reward function in such a way that minimizes different forms of optimal behavior vs the expert's demonstrated behavior (Nguyen, Low, & Jaillet, 2015). As such, Abbeel and Ng (2004) advised a projection algorithm to find a reward function by matching the feature expectations generated by choosing an optimal policy with the feature expectations obtained by the expert's observed trajectories. Called Apprenticeship Learning (AL), this algorithm provided a base for further investigations and studies of the IRL. Most of the AL methods include several specific items such as Feature Expectation Matching (FEM) (Abbeel & Ng, 2004), Max-Margin Planning (MMP) (Ratliff et al., 2006a), feature boosting (Ratliff et al., 2006b; Bagnell, Chestnutt, Bradley, & Ratliff, 2007; Ratliff, Silver, & Bagnell, 2009; Zucker et al., 2011; Silver, Bagnell, & Stentz, 2010), Policy Matching (PM) (Neu & Szepesvári, 2007), Bayesian IRL (BIRL) (Ramachandran & Amir, 2007), hierarchical formulation (Kolter, Abbeel, & Ng, 2008), Multiplicative Weight Apprenticeship Learning algorithm (MWAL) (Syed & Schapire, 2008; Syed, Bowling, & Schapire, 2008), Maximum-Entropy IRL (MaxEntIRL) (Ziebart, Maas, Bagnell, & Dey, 2008), Learning to Search (LEARCH) (Ratliff et al., 2009; Paraschos et al., 2013), minimizing the sum of squares error (SSE) (Mombaur, Truong, & Laumond, 2010), Feature construction IRL (FIRL) (Levine et al., 2010), Maximum Likelihood IRL (MLIRL) (Babes, Marivate, Subramanian, & Littman, 2011), Relative entropy IRL (REIRL) (Boularias, Kober, & Peters, 2011), Partially Observable Markov Decision Process (POMDP) (Choi and Kim, 2011), Gaussian Process IRL (GPIRL) (Levine et al., 2011), Structured Classification IRL (SCIRL) (Klein, Geist, Piot, & Pietquin, 2012), Continuous Inverse Optimal Control (CIOC) (Levine & Koltun, 2012), Linearly-solvable MDPs (LMDPs) (Dvijotham & Todorov, 2010), Path Integrals IRL (PI-IRL) (Kalakrishnan, Pastor, Righetti, & Schaal, 2013; Aghasadeghi & Bretl, 2011b), IOC Differential Dynamic Programming (IOCDDP) (Park & Levine, 2013), Maximum Entropy Semi-Supervised IRL (MESS IRL) (Audiffren, Valko, Lazaric, &

Ghavamzadeh, 2015), Bayesian Nonparametric IRL(BNIRL) (Choi & Kim, 2013; Michini, Walsh, Agha-Mohammadi, & How, 2015), Deep Maximum Entropy IRL(DMaxEntIRL) (Wulfmeier, Ondruska, & Posner, 2015), Deep Inverse Reinforcement Learning (Deep AP: Deep Apprenticeship Q-Network (DAQN) and Deep Apprenticeship Reward Network (DARN) (Bogdanovic, Markovikj, Denil, & De Freitas, 2015), Deep Gaussian Process IRL (DGP-IRL) (Jin, Damianou, Abbeel, and Spanos, 2015), guided cost learning/deep IOC (Finn et al., 2016a), Generative Adversarial Imitation Learning (GAIL) (Ho & Ermon, 2016), Generative Adversarial Maximum Entropy (GAMaxEntIRL) (Finn et al., 2016b), Generative Adversarial Network (OptionGAN) (Henderson et al., 2018), Policy Gradient IRL (PGIRL) (Pirotta & Restelli, 2016), inverse KKT (Englert, Vien, & Toussaint, 2017), and Compatible Reward IRL (CRIRL) (Metelli, Pirotta, & Restelli, 2017) – the list is not exhaustive. When it comes to the demonstration, most of the mentioned methods assume that the expert's demonstrated trajectories are generated by a single reward function. In the meantime, there are cases (Babes et al., 2011; Choi & Kim, 2011) where the trajectories are generated by multiple reward functions but, at the same time, each trajectory is still assumed to be produced by a single reward function (Nguyen et al., 2015). Regarding the problem modeling, one may refer to the model-based approaches such as those proposed by Abbeel and Ng (2004), Ratliff et al. (2006a), Ziebart et al. (2008), Silver et al. (2010), Ziebart (2010), Levine et al. (2011), Levine and Koltun (2012), Hadfield-Menell, Russell, Abbeel, and Dragan (2016), and Finn et al. (2016b) and the model-free approaches including, but not limited to, the formulations presented by Boularias et al. (2011) and Kalakrishnan et al. (2013).

In addition to the approaches outlined above, there are other important areas associated with the IOC or IRL that worth reviewing. For example, Constrained IOC/IRL (Harder & Wachsmuth, 2019; Doerr, Ratliff, Bohg, Toussaint, & Schaal, 2015; Pauwels et al., 2014; Gaurav & Ziebart, 2019; Menner, Worsnop, & Zeilinger, 2018, 2019), inverse optimization (Ahuja & Orlin, 2001; Heuberger, 2004; Iyengar and Kang, 2005) and inverse sub-Riemannian geometric problem (Maslovskaya, 2018), as well as, two Policy Gradient (PG) methods, namely Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO) (Schulman, Levine, Abbeel, Jordan, & Moritz, 2015; Schulman et al., 2017) have been recently used for Generative Adversarial Imitation Learning (GAIL) (Ho & Ermon, 2016; Henderson et al., 2018).

One of the most practical topics associated with IOC is the learning of human behaviors such as walking, locomotion, and biological movements (Arechavaleta, Laumond, Hicheur, & Berthoz, 2008; Mombaur et al., 2010; Li et al., 2011; Puydupin-Jamin, Johnson, & Bretl, 2012; Aghasadeghi, 2015; Clever & Mombaur, 2016), human skill learning (Ghalamzan, Amir, Paxton, Hager, & Bascetta, 2015), and application of IRL to games (Uchibe, 2018; Hessel et al., 2018; Tucker, Gleave, & Russell, 2018) in the context of IOC problem to learn the cost functions and optimality criteria of the human behavior.

Many surveys, reviews, and papers have reported on LfD, IL, and IOC problems in several areas such as level set methods for inverse problems and optimal design (Burger & Osher, 2005), IRL theory, algorithms, techniques, challenges, and recent advances (Ng & Russell, 2000; Abbeel, Coates, & Ng, 2010; Zhifei & Joo, 2012; Gao, Peters, Tsourdos, Zhifei, & Joo, 2012; Arora & Doshi, 2018), IL and LfD (Argall, Chernova, Veloso, & Browning, 2009; Bandera, Rodriguez, Molina-Tanco, & Bandera, 2012; Billard & Grollman, 2013; Hussein, Gaber, Elyan, & Jayne, 2017; Zhu & Hu, 2018), Bayesian reinforcement learning (Ghavamzadeh, Mannor, Pineau, & Tamar, 2015), deep apprenticeship learning (Markovikj, 2014; Arulkumaran, Deisenroth, Brundage, & Bharath, 2017), and generative adversarial imitation learning (Osa et al., 2018). Also, a few titles have been published on the IOC. Neittaanmäki, Rudnicki, Rudnicki, and Savini (1996) published a book on the inverse problems and optimal design in the field of electricity and magnetism, covering several areas such as electrical machines, high-voltage engineering, nuclear magnetic resonance spectrography, electron optics, plasma techniques, etc. Widrow and Walach (2008) wrote a book on adaptive IOC for signal processing. Sanchez and Ornelas-Tellez (2017) contributed a title where discrete-time IOC was explained and the passivity and CLF approaches were described and demonstrated on practical examples.

3. General formation of IOC

Consider the definition of LfD as described in Section 1. Given a system dynamics $\dot{x} = f(t, x(t), u(t))$, where, $t \in [0\,T]$, $x \in \mathbb{R}^n \in X$ and $u \in \mathbb{R}^m \in U$ are time, state, and control variables, respectively, the initial and final states $x(0) = x_s$ and, $x(t_f) = x_f$, and a demonstration trajectory dataset $D = \{(x_1, u_1), \ldots, (x_n, x_n)\}$, an agent tries to find an originally unknown Lagrangian function $l(t,\ x, u) \in [0\ T] \times \mathbb{R}^n \times \mathbb{R}^m$ to generate a similar path trajectory to the dataset D by satisfying the related constraints and the initial and final conditions. Formally, this problem can be stated as below:

$$l = LfD(x,u)\ s.t.\ \begin{array}{c} \min\limits_{x,u} V(t,x) = \int_{t_0}^{t_f} l(t, x(t), u(t))dt \\ \dot{x} = f(t, x(t), u(t)) \\ x(0) = x_s,\ x\left(t_f\right) = x_f \end{array} \tag{1}$$

In most cases, either the Lagrangian is in the form of features or one must approximate it by a linear weighted feature-based function. Based on this idea, a primal problem of feature-based learning from demonstration is stated as:

$$w = LfD(x,u)\ s.t.\ \begin{array}{c} \min\limits_{u} V(t,x) = \int_{t_0}^{t_f} w^T \phi(t, x(t), u(t))dt \\ \dot{x} = f(t, x(t), u(t)) \\ x(0) = x_s,\ x\left(t_f\right) = x_f \\ D = \left\{x_i^E, u_i^E\right\}_{i=1,\ldots,m} \end{array} \tag{2}$$

where t is time, x and u are state and control variables, respectively, and $\phi(t,\ x(t),\ u(t))$ is the feature vector or basis function. In this case, the goal of the learning from demonstration problem is to find the weighting matrix w by minimizing the objective function $V(t,\ x)$ w.r.t. the control variable u subject to the system dynamics $\dot{x} = f(t, x(t), u(t))$ given the initial and final values $x(0) = x_s$ and $x(t_f) = x_f$, respectively, and the data trajectories set $D = \{x_i^E, u_i^E\}_{i=1,\ldots,m}$ that is observed from a demonstration by an expert doing a control or learning task.

Given the system $\dot{x} = f(t, x(t), u(t))$ and the objective function $V(t,\ x)$ constructed by the Lagrangian function $l(t, x,\ u)$, the optimal control problem is defined to determine an optimal control $u^*(t)$ whereby the optimality of the problem is evaluated in a long-term scheme as follows:

$$u^* = OCP(t, x, f, g, h)\ \begin{array}{c} \underset{u}{minimiz}\ V(x(t), u(.), t) \triangleq \Psi(x(T)) \\ + \int_t^T l(t, x(\tau), u(\tau), \tau)d\tau \\ subject\ to\ \dot{x} = f(t, x(t), u(t));\ x(t_0) = x_0 \\ g_i(x,u) \le 0,\ i = 1, \ldots, N_{ineq} \\ h_j(x,u) = 0,\ j = 1, \ldots, N_{eq} \end{array} \tag{3}$$

where t_0 and t_f are the initial and final times, respectively, Ψ and l are scalar functions. t_f may be specified or "free," depending on the problem statement. N_{ineq} and N_{eq} are the numbers of inequalities $g(x,\ u)\ \le\ 0$ and equalities $h(x, u)\ =\ 0$, respectively.

The problem expressed by Eq. (3) is a primal formulation of a constrained optimal control. In most cases, to solve an optimal control problem, especially for non-linear and non-LQR systems,

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

the corresponding dual problem must be established. By defining the Hamiltonian as $H(t, x, u, w, v, \lambda) = l(t, x, u) + pf(t, x, u) + vg(x, u) + \lambda h(x, u)$, minimizing it w.r.t. the variable u assuming $\frac{\partial H}{\partial u} = 0$ to find an optimal control law $u^*(t)$, and then substituting it into the HJB equation $HJB: \frac{\partial p}{\partial t} + H(t, x, u^*(t), p, v, \lambda) = 0$, the dual problem for the optimal control problem Eq. (3) can be stated as follows:

$$(u^*, v, \lambda, p) = OCP(t, x, f, g, h) \quad \begin{array}{c} u^*(t) = argmin_u H(t, x, u, p, v, \lambda) \\ H = l(t, x, u) + pf(t, x, u) + v^T g(x, u) + \lambda^T h(x, u) \\ \frac{\partial p}{\partial t} + H = 0;\ \frac{\partial H}{\partial p} - f(t, x, u^*) = 0;\ \frac{\partial H}{\partial x} - p = 0 \\ \frac{\partial H}{\partial v} - g(x, u) = 0;\ \frac{\partial H}{\partial v} - h(x, u) = 0 \\ v^T g(x, u^*) = 0, \lambda^T h(x, u^*) = 0 \\ v \geq 0, \lambda \geq 0, p \geq 0, l \geq 0 \end{array} \tag{4}$$

where v and λ are the Lagrange multipliers associated with the inequality constraint $g(x, u) \leq 0$ and equality $h(x, u) = 0$, respectively, and p is the Hamiltonian multiplier. The Lagrange multipliers v and λ are also called the dual variables or Lagrange multiplier vectors associated with the problem. Finding a solution to such a problem is generally difficult but it can be solved by simplifying assumptions and using optimization and IOC methods.

To solving the optimal control problem for nonlinear discrete-time systems with known mathematical models, which does not require an initially stable policy, Heuristic Dynamic Programming (Werbos, 1992), is proposed to obtain the optimal control law (Wang et al., 2009). The problem IOC can be stated as

$$(r, V) = IOC(x, u)\ s.t. \quad \begin{array}{c} \min_u V(x_k) = \sum_{i=k}^{+\infty} \gamma^{i-k}\, r(x_i, u_i) \\ x(k+1) = f(x_k) + g(x_k)u_k \\ x(0) = x_s,\ x\left(t_f\right) = x_f \end{array} \tag{5}$$

Where, $r(x_k, u_k) = x_k^T Q x_k + u_k^T R u_k$, $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{n\times n}$ are positive definite matrices. Starting from $V(x_0) = 0$, the state-dependent cost function can be rewritten as $V(x_k) = r(x_k, u_k) + \sum_{i=k+1}^{+\infty} \gamma^{i-(k+1)}\, r(x_i, u_i) = r(x_k, u_k) + \gamma V(x_{k+1})$. Now, instead of evaluating the infinite sum as $V(x_k) = \sum_{i=k}^{+\infty} \gamma^{i-k}\, r(x_i, u_i)$, one can solve the differential equation to obtain the value of using a current policy $u_k = h(x_k)$. This is a nonlinear Lyapunov equation known as the Bellman equation that is the first key concept in developing reinforcement learning techniques (Lewis & Liu, 2013). Then, the discrete-time Hamiltonian function can be defined as $H(x_k,\ u_k, \Delta V_k) = r(x_k, u_k) + \gamma V(x_{k+1}) - V(x_k)$. Where, $\Delta V_k = \gamma V(x_{k+1}) - V(x_k)$. Now, by assuming the gradient of hamiltonian $\frac{\partial}{\partial u_k} H(x_k,\ u_k, \Delta V_k) = \frac{\partial r(x_k, u_k)}{\partial u_k} + \frac{\partial x_{k+1}}{\partial u_k} \frac{\partial V^*(x_{k+1})}{\partial x_{k+1}} = 0$, the optimal policy u_k^* can be resulted as $u^*(x_k) = -\frac{\gamma}{2} R^{-1} g(x_k)^T \frac{\partial V(x_{k+1})}{\partial x_{k+1}}$ and the corresponded value function $V^*(x_k)$ is yielded as $V^*(x_k) = x_k^T Q x_k + \frac{\gamma^2}{4} \frac{\partial V(x_{k+1})}{\partial x_{k+1}} g(x_k) R^{-1} g(x_k)^T \frac{\partial V(x_{k+1})}{\partial x_{k+1}}$. This equation reduces to the Riccati equation in the LQR case, which can be efficiently solved but generally the HJB cannot be solved exactly for nonlinear cases (Al-Tamimi et al., 2008) and often the solution of such problems needs to new assumptions such as the value function is a positive definite quadratic function as $V(x) = x^T P x$, where P is a symmetric positive definite matrix, or the system is passive with the passivity condition $\Delta V_k \leq y_k^T u_k$, where, y_k is the system output, or the state can be approximated by neural networks (Sanchez & Ornelas-Tellez, 2017), or the value function is approximated by heuristic methods in the context of approximate dynamic programming and reinforcement learning (Al-Tamimi et al., 2008; Wang et al., 2009; Lewis & Liu, 2013).

The solution of the HJB problem exists only for the linear regulator problem, for which it is particularly well-suited Anderson & Moore, 2007). Hence, to avoid the HJB equation solution, the IOC approach was proposed (Moylan & Anderson, 1973; Sepulchre, Jankovic, & Kokotovic, 2012). This approach begins with synthesizing a stabilizing feedback control law and follows to establish that this control law optimizes a cost function (Ornelas et al., 2010). In other words, mathematically speaking, an IOC is a set of differential equations describing the shape of an objective function minimized w.r.t a control law given the initial state and state/control-sampled data. i.e. $D = \{(x_i,\ u_i)\}_{i=1}^n$. Assuming that the data D maps to a stabilizing control law K^D, one can find the objective function for a given stable dynamic system. For example, given a stabilizable system $\dot{x} = Ax(t) + Bu(t)$ and a constant stabilizing feedback control law $u(t) = Kx(t)$, the first classic IOC (also known as Kalman IOC problem (Kalman, 1964) is stated as follows: 1. existence: determine the necessary and sufficient conditions on matrices A, B and K, such that K is an optimal control law for some cost function $l(x, u) = x^T(t)Qx(t) + u^T(t)Ru(t)$, and 2. solution: determine matrices R and Q corresponding to the same K (Li et al., 2018).

$$(Q, P, R) = IOC(K,\ A, B, x, u): \quad \begin{array}{c} \min_u V(x_0, T, u) = \lim_{t\to\infty} \int_{t_0}^{T} L(t, x, u)dt \\ \dot{x} = Ax(t) + Bu(t) \\ u(t) = -kx(t) \\ L(t, x,\ u) = \frac{1}{2} x^T Q x + P^T x u + \frac{1}{2} R u^T u \end{array} \tag{6}$$

In recent years, the IOC problem has been introduced into the concept of the IRL or AL while there are some differences between these approaches, as shown in Table 1. The same procedure can be applied for the IRL problems with the exception that the control policy is not assumed as a stabilizing control law but it is supposed that the system behavior is optimal and its effect can be accessed by demonstration or sampling data of state and control variables in a finite time. This is the key difference between IOC and IRL.

Supposing finite sets of states S and actions A, policy $\pi \in \Pi: S \to A$ where Π is the set of all stationary stochastic policies that take actions in A given states in S as the successor states that are drawn from the dynamics model $P(s'|s, a)$ and an immediate reward (or expected immediate reward with discounting factor γ) received after transitioning from state s to state s', due to action a, a reinforcement learning problem is formed in the framework of a Markov decision process $MDP(S, A, P, R, \gamma)$ to find a policy π that specifies the action $a = \pi(s)$ for an agent that chooses this policy in state s. The goal is to select a policy π that maximizes some cumulative function of the rewards – typically the expected discounted sum over an infinite horizon $\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)$, where s_t and a_t are the state and action at time t, respectively. For a policy π, the value function $V_\pi: S \to R$ gives the value of a state as the long-term expected cumulative reward incurred from the state by following π. The value function of a policy π for a given state s is written as $V^\pi(s) = E[\sum_{s'} P(s'|s, a)(R(s, a) + \gamma V^\pi(s'))]$. The expected discounted sum of rewards by starting in state s and opting for action a by taking the policy π is given by the Q-function $Q^\pi(s, a) = E[R(s, a) + \gamma \sum_{s'} P(s'|s, a) V^\pi(s')]$. Using this function, the optimal policy can be obtained as $\pi^* = \arg\max_{a\in A} Q^*(s, a)$. In a reinforcement learning problem, the goal is to find an optimal policy that maximizes the state value function $V^\pi(s)$ or state-action Q-function $Q^\pi(s, a)$ given an MDP. An IRL, however, looks to find or reconstruct the reward function R given an optimal policy π and MDP\R. Typically, IRL problem can be formulated in the framework

Table 1
Inverse Reinforcement Learning (IRL) versus Classic Inverse Optimal Control (IOC).

Items	Modern IOC (IRL)	Classic IOC
Space	State-space $S \subset \mathbb{R}^n$، action set $A \subset \mathbb{R}^m$	State-space $X \subset \mathbb{R}^n$, control space $U \subset \mathbb{R}^m$
Variables	State $s \in S$, action $a \in A$	State $x \in X$, control $u \in U$
Objective function	Reward function $R(s, a)$/cost function $C(s, a)$ or cumulative discounted function $V(s_t, a_t) = \max_{\pi} E[\sum_t \gamma^t R(s_t, a_t)\|\pi]$	Cost function or performance index $J(t, x, u)$ based on a CLF or Lagrangian function $l(x, u)$
Control function	Optimal Policy $\pi(s, a)$	Stabilizing Control law $k(x)$
Given demonstration	State-action trajectory set $D = \{(s_1, a_1), \ldots, (s_n, a_n)\}$	State-control trajectory set $D = \{(x_1, u_1), \ldots, (x_n, x_n)\}$
Model	Model-free or model-based as Markov Decision Process (MDP): $s_{t+1} = M(s_t,\ a_t, T, R, \gamma)$	Differential equations: $\dot{x} = f(x, u)$
Problem formulation	$R = IRL(S, A) \quad \begin{aligned} V(s_t, a_t) &= \max_{\pi} E[\textstyle\sum_t \gamma^t R(s_t, a_t)\|\pi] \\ s.t.\ s_{t+1} &= M(S, A,\ T, \gamma)\backslash R \\ s(0) &= s_0 \\ s(t_f) &= s_{t_f} \end{aligned}$	$l = IOC(x, u) \quad \begin{aligned} J(x, u) &= \min_{x} \textstyle\int_{t_0}^{t_f} l(x, u)dt \\ s.t.\ \dot{x}(t) &= f(x, u) \\ x(0) &= x_0 \\ x(t_f) &= x_{t_f} \end{aligned}$

of MDP, as follows:

$$R = IRL(S, A) \quad \begin{aligned} V(s_t, a_t) &= \max_{\pi} E\left[\sum_t \gamma^t R(s_t, a_t)|\pi\right] \\ s.t.\ s_{t+1} &= M(S, A,\ T, \gamma)\backslash R \\ s(0) &= s_0 \\ s\left(t_f\right) &= s_{t_f} \\ D &= \{(s_1, a_1), \ldots, (s_n, a_n)\} \end{aligned} \tag{7}$$

where $s(0) = s_0$ and $s(t_f) = s_{t_f}$ are the given initial and final states, respectively, D is a given demonstrated trajectory that is used by an agent to find the reward function R by maximizing the expected cumulative reward function $E[\sum_t \gamma^t R(s_t, a_t)|\pi]$. In most cases, an expert policy is used instead of an optimal policy. This is while the expert policy is often not available and the real effects of a policy that are measurable, observable, or interpretable are used alternatively. For example, the demonstration of an expert's action when doing a task or measurement of the action outputs by sensors or captured images of the action helps us to extract the expert's behavior policy. Taking the reward function as $R(s, a) = w^T\phi(s, a)$, where $w \in \mathbb{R}^n$ is a weighting vector and, $\phi : (S, A) \to \mathbb{R}^n$ is the basis function denoting the features and constraints of the system and the environment, the problem of deriving a reward function from the observed behavior is referred to as IRL (Ng & Russell, 2000), which is described in the next chapters.

4. Categorization of the IOC methods

Based on a widely accepted approach by Torabi et al. (2018), the IL/LfD problems have been categorized into two branches, namely BC and IOC, as shown in Fig. 1 However, a combination of BC with IRL has been also proposed recently by Metelli et al. (2017).

Given a finite state space with sets of states $x \in X$ and controls $u \in A$ for a control system or states $s \in S$ and actions $a \in A$ corresponding to the learning problem with the demonstrated data $D_c = \{u_k, x_k\}_{k=1}^n$ for the control system or $D_L = \{a_k, s_k\}_{k=1}^n$ corresponding to the learning process, the goal of an IL/LfD is to find a control law $k(x(t))$ or policy $\pi(s)$ in such a way that the corresponding controller/agent can follow/generate the trajectory ξ^D related to the dataset D. To do this, one needs a cost/reward function $c(x, u)$ or $r(s, a)$ to evaluate whether a trajectory ξ generated by the policy π is similar to the demonstrated trajectory ξ^D or not. Formulation of the problem into the framework of a supervised learning approach such as regression method to find an optimal policy leads to a BC problem, while the formulation of the initial problem to find the cost/reward function that is being stabilized or optimized by the agent's policy results in an IOC problem.

A classic IOC problem addresses the recovery of the cost/reward function given the state and/or control samples of a stable system, while the modern IOC (IRL) aims to find a reward function under which the optimal policy matches with the expert's demonstrations (Levine et al., 2011). Generally, it is implicitly accepted that the expert's behavior is presumably optimal in IRL, although in reality, as Kalman (1964) stipulates, the "optimality does not imply stability". Most of the classic IOC methods benefit the stabilization assumption, However, under certain conditions (*i.e.* stabilizability and detectability), the stability of the closed-loop system can be guaranteed by choosing an appropriate CLF as the optimal control value function (Xi & Li, 2019). The most popular classic IOC methods are shown in Fig. 1 Generally, the following definitions are of help for understanding the BC, classic IOC, and modern IOC (IRL).

- **BC:** Behavior Cloning problem is to learn or find an optimal policy by matching(supervised learning) directly the state/control trajectory generated by the policy and the given demonstration data (Osa et al., 2018; Bain and Sammut, 1996).
- **Classic IOC:** Inverse Optimal Control problem is to find or reconstruct the cost/reward function given a stabilizing control law(Demonstration data is the result of a stable system).
- **Modern IOC (IRL):** Inverse Reinforcement Learning problem is to find or reconstruct the cost/reward function given a demonstration data generated by an optimal policy(Demonstration data is the result of optimal behavior of the system).

There are three approaches for recovering the cost/reward function under the IRL paradigm, as shown in Fig. 2 The first approach refers to the direct methods that attempt to learn the policy by optimizing a loss function that measures the deviation between the expert's policy and the policy chosen in the framework of the supervised learning. The second approach focuses on the indirect methods where the expert is assumed to be acting optimally in the environment modeled as an MDP (Neu & Szepesvári, 2007). An example of the indirect approach is the first IRL technique proposed by Ng and Russell (2000) as a reward learning method that derived a reward function from observed behavior by assuming the optimality. In this method, the optimal policy is indirectly obtained after finding the reward function. Proposed by Abbeel and Ng (2004), AL is a direct IRL approach in which the algorithm finds a directly optimal policy (and/or value function) performing similar to the expert's with unknown reward function and approximated by parametrizing the weighted feature-based reward function with unknown weights and known features. Neu and Szepesvári (2007) combined these two approaches by using a loss function that minimizes the deviation from the expert's policy, like

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

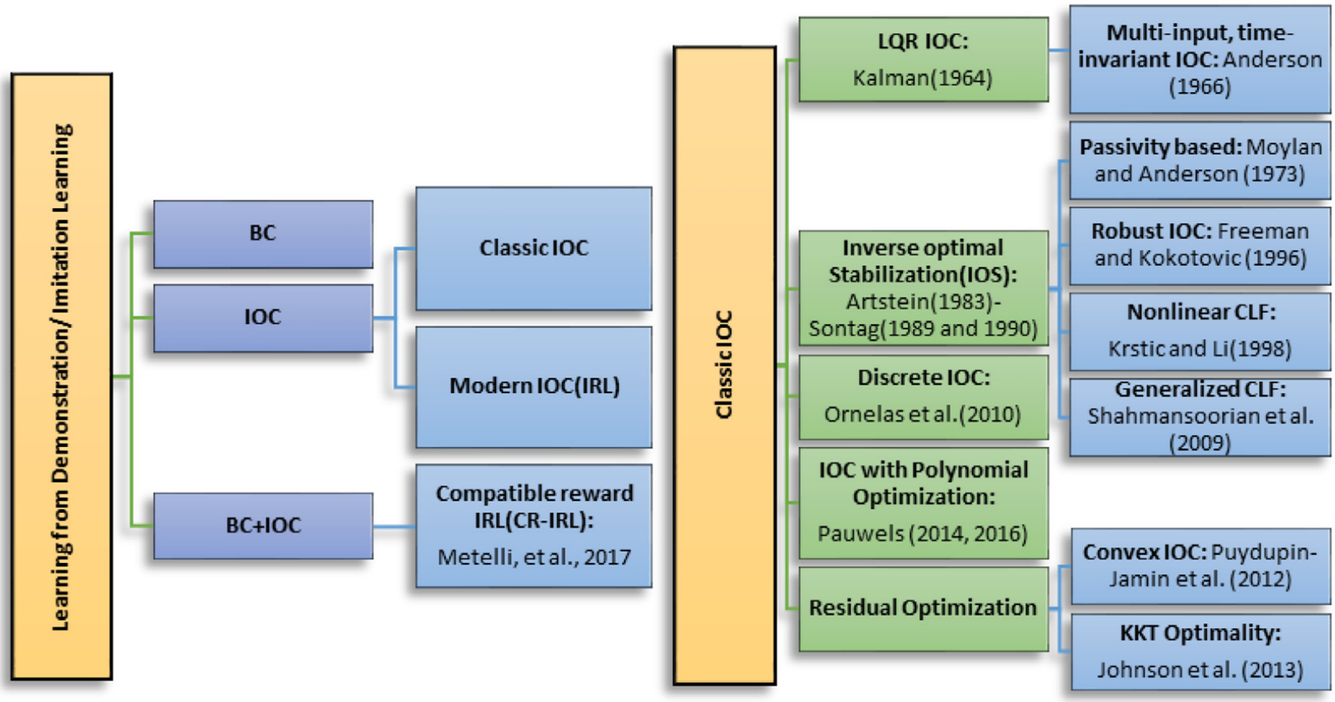
Learning from Demonstration/Imitation Learning
BC
IOC
BC+IOC
Classic IOC
Modern IOC(IRL)
Compatible reward IRL(CR-IRL): Metelli, et al., 2017
Classic IOC
LQR IOC: Kalman(1964)
Inverse optimal Stabilization(IOS): Artstein(1983)-Sontag(1989 and 1990)
Discrete IOC: Ornelas et al.(2010)
IOC with Polynomial Optimization: Pauwels (2014, 2016)
Residual Optimization
Multi-input, time-invariant IOC: Anderson (1966)
Passivity based: Moylan and Anderson (1973)
Robust IOC: Freeman and Kokotovic (1996)
Nonlinear CLF: Krstic and Li(1998)
Generalized CLF: Shahmansoorian et al. (2009)
Convex IOC: Puydupin-Jamin et al. (2012)
KKT Optimality: Johnson et al. (2013)

Fig. 1. Left: Classification of IL/LfD methods into the BC and IOC problems, and the IOC method further into classic and modern categories. Right: Classic IOC methods.

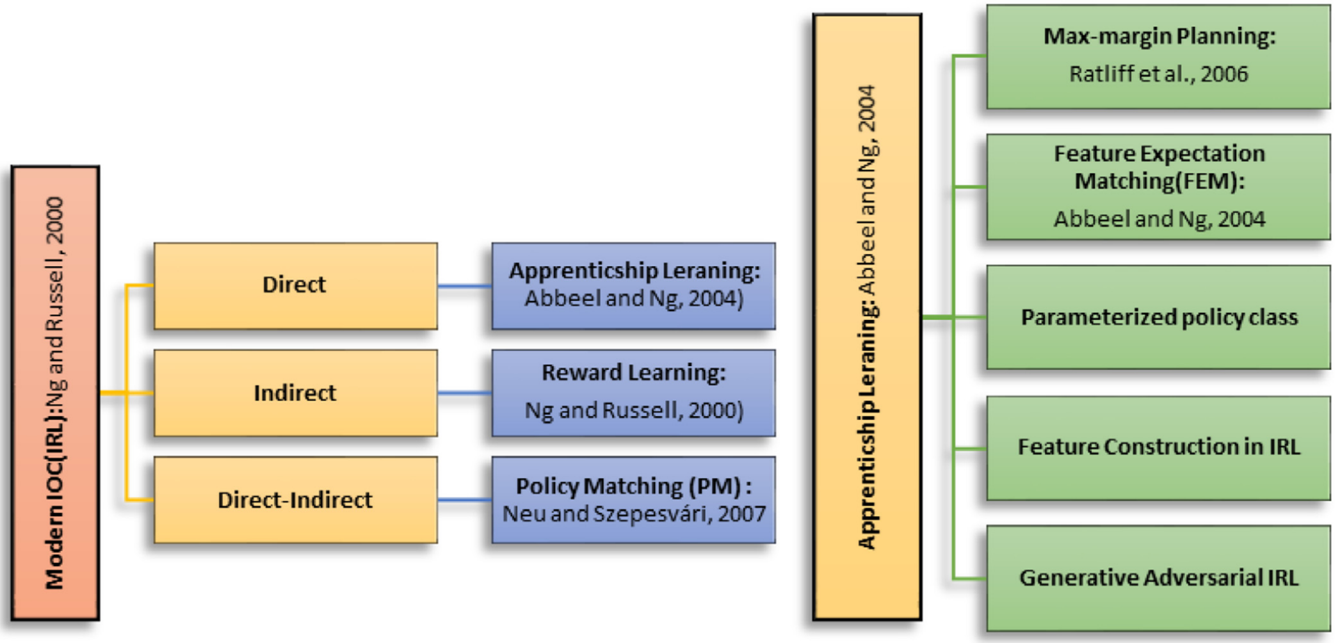
Modern IOC(IRL):Ng and Russell, 2000
Direct
Indirect
Direct-Indirect
Apprenticship Leraning: Abbeel and Ng, 2004)
Reward Learning: Ng and Russell, 2000)
Policy Matching (PM): Neu and Szepesvári, 2007
Apprenticship Leraning: Abbeel and Ng, 2004
Max-margin Planning: Ratliff et al., 2006
Feature Expectation Matching(FEM): Abbeel and Ng, 2004
Parameterized policy class
Feature Construction in IRL
Generative Adversarial IRL

Fig. 2. Modern IOC techniques (i.e. IRL).

in supervised learning, with the policy itself obtained by tuning a reward function and solving the resulting MDP, instead of finding the parameters of the policy.

The AL has been classified under five main categories, namely Max-Margin Planning (MMP), FEM, parameterized policy class, feature construction in IRL, and generative adversarial IRL (see Figs. 2 and 3). These five categories have been covered in particular IRL works, as shown in Fig. 3 and Fig. 4.

5. IOC problems

5.1. Classic IOC

Kalman Problem: The first problem in the scope of IOC was introduced and formulated by Kalman (1964). Considering the system $\dot{x} = Ax(t) + Bu(t)$ where, x, u, A, and B are state and control variables and state and control matrix coefficients, respectively,

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

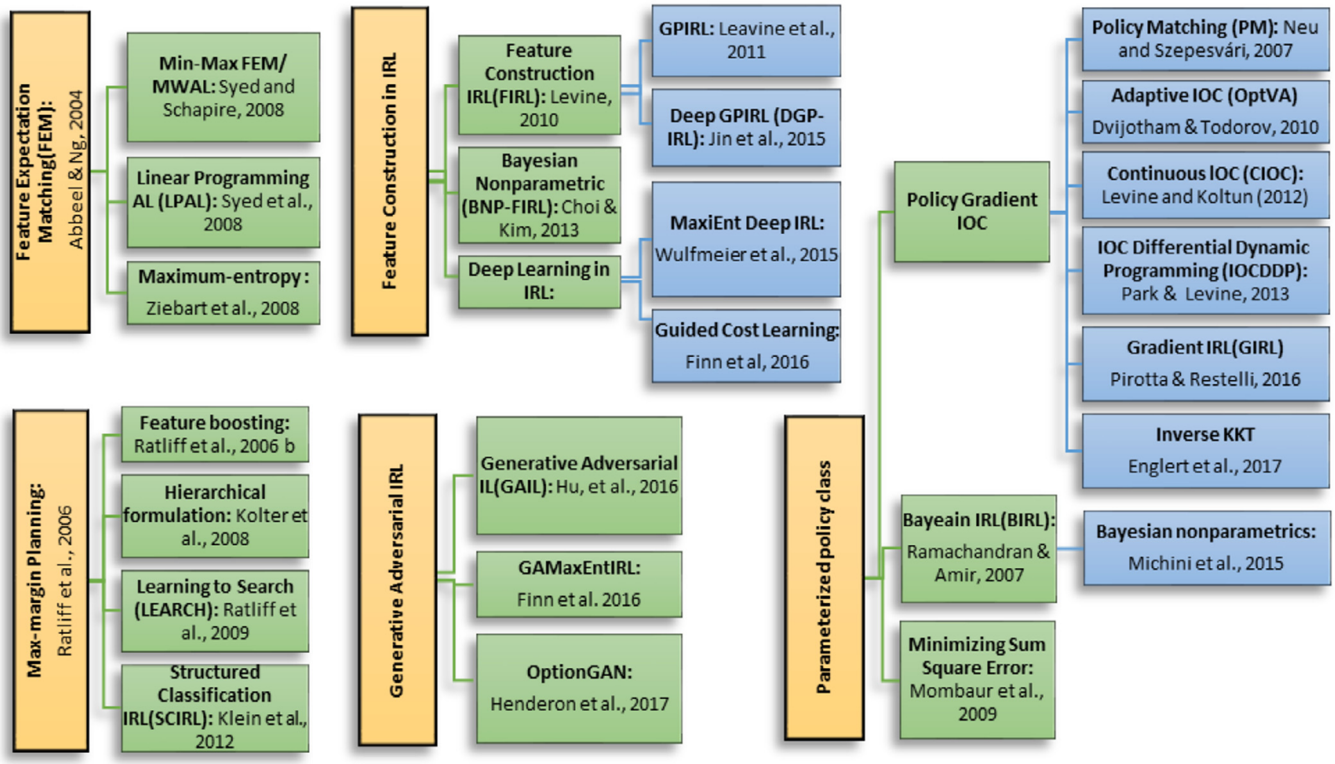
Feature Expectation Matching(FEM): Abbeel & Ng, 2004
Min-Max FEM/ MWAL: Syed and Schapire, 2008
Linear Programming AL (LPAL): Syed et al., 2008
Maximum-entropy: Ziebart et al., 2008
Feature Construction in IRL
Feature Construction IRL(FIRL): Levine, 2010
Bayesian Nonparametric (BNP-FIRL): Choi & Kim, 2013
Deep Learning in IRL:
GPIRL: Leavine et al., 2011
Deep GPIRL (DGP-IRL): Jin et al., 2015
MaxiEnt Deep IRL: Wulfmeier et al., 2015
Guided Cost Learning: Finn et al, 2016
Policy Gradient IOC
Policy Matching (PM): Neu and Szepesvári, 2007
Adaptive IOC (OptVA) Dvijotham & Todorov, 2010
Continuous IOC (CIOC): Levine and Koltun (2012)
IOC Differential Dynamic Programming (IOCDDP): Park & Levine, 2013
Gradient IRL (GIRL) Pirotta & Restelli, 2016
Inverse KKT Englert et al., 2017
Max-margin Planning: Ratliff et al., 2006
Feature boosting: Ratliff et al., 2006 b
Hierarchical formulation: Kolter et al., 2008
Learning to Search (LEARCH): Ratliff et al., 2009
Structured Classification IRL(SCIRL): Klein et al., 2012
Generative Adversarial IRL
Generative Adversarial IL(GAIL): Hu, et al., 2016
GAMaxEntIRL: Finn et al. 2016
OptionGAN: Henderon et al., 2017
Parameterized policy class
Bayeain IRL(BIRL): Ramachandran & Amir, 2007
Bayesian nonparametrics: Michini et al., 2015
Minimizing Sum Square Error: Mombaur et al., 2009

Fig. 3. AL methods.

and given the stable control law $u(t) = -kx(t)$ and performance index $V(x_0,\ \infty, u) = \lim_{t\to\infty} \int_{t_0}^{t} L(x, u)dt$, the goal of this problem is to find P, Q, and R given a completely controllable constant linear plant A, B, k, and the control law $u(t)$ such that the control law minimizes the performance index V. Using this definition, the IOC is formed as follows:

$$(k, Q, P, R) = IOC(x, u,\ A,\ B) : \begin{array}{c} \min\limits_{u} V(x_0, T, u) = \lim\limits_{t\to\infty} \int_{t_0}^{T} L(t, x, u)dt \\ \dot{x} = Ax(t) + Bu(t) \\ u(t) = -kx(t) \\ L(t, x,\ u) = \frac{1}{2}x^TQx + P^Txu + \frac{1}{2}Ru^Tu \end{array} \tag{8}$$

where, Q, P, and R are the weighting matrices. Hereinafter, $IOC(x,\ u)$ denotes the IOC problem with the inputs x and u. Kalman tried to solve this problem using frequency-domain methods. This problem provided a theoretical basis for further works in the field of IOC.

Passivity-based IOC problem: Moylan and Anderson (1973) proposed an IOC problem for the non-linear regulator theory. For the system $\dot{x} = f(x(t)) + gu(t)$ with the asymptotically stable control law, the IOC problem was structured as below:

$$l = IOC(x, u) : \begin{array}{c} \min\limits_{u} V(x(t_0), u(.), t_0,\ T) = \lim\limits_{t\to\infty} \int\limits_{t_0}^{T} \left(l(x(t)) + u(t)^T u(t)\right)dt \\ \dot{x} = f(x(t)) + gu(t) \\ u(t) = -k(x(t)) \\ \int_{t_0}^{T} (u + k(x))^T (u + k(x))dt \geq \int_{t_0}^{T} u^T udt \end{array} \tag{9}$$

where the term $\int_{t_0}^{T} (u + k(x))^T (u + k(x))dt \geq \int_{t_0}^{T} u^T udt$ is a passivity condition. By forming the Hamiltonian function $H(x(t), u(t), \frac{\partial \phi}{\partial x}) = l(x(t)) + u(t)^T u(t) + \frac{\partial \phi(t,x(t),T)}{\partial x}(f(x(t)) + gu(t))$, and minimizing it w.r.t. u, the control law could be obtained as $u(t) = -k(x(t)) = -\frac{1}{2}\, g(t)^T \frac{\partial \phi^T(t,x(t),T)}{\partial x}$, and the function $l(x(t))$ was reconstructed as

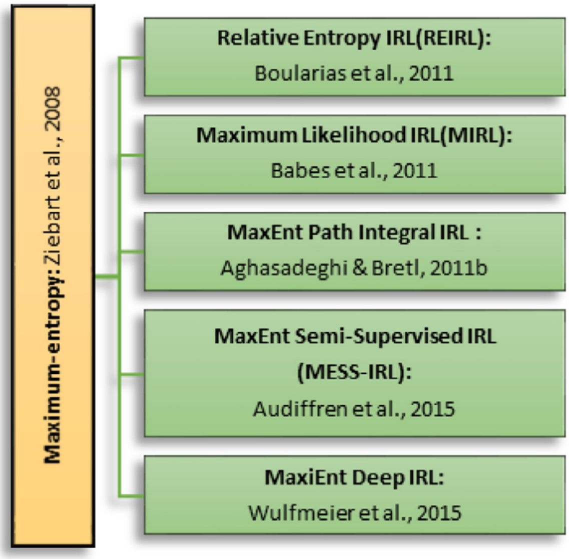
Maximum-entropy: Ziebart et al., 2008
Relative Entropy IRL(REIRL): Boularias et al., 2011
Maximum Likelihood IRL(MIRL): Babes et al., 2011
MaxEnt Path Integral IRL : Aghasadeghi & Bretl, 2011b
MaxEnt Semi-Supervised IRL (MESS-IRL): Audiffren et al., 2015
MaxiEnt Deep IRL: Wulfmeier et al., 2015

Fig. 4. Maximum-entropy methods on IRL.

$l(x(t)) = -\frac{\partial \phi}{\partial x}(f(x(t)) + \frac{1}{4}gg^T \frac{\partial \phi^T(t,x(t))}{\partial x})$, where $\phi^T(t,\ x(t),\ T)$ is the solution of $\frac{\partial \phi(t,x(t))}{\partial t} + H(x(t),\ \frac{\partial \phi}{\partial x}) = 0$ and can be obtained by assuming the mentioned passivity condition and other mathematical assumptions that are expressed in detail in the mentioned reference.

Constructive CLF: The CLF is used as a powerful tool to determine whether a dynamic system is stable (more specifically, asymptotically stable). The so-called Artstein-Sontag theorem (Artstein, 1983; Sontag, 1989) states that a dynamic system has a differentiable CLF if and only if there exists regular sta-

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

bilizing feedback (Lin & Sontag, 1991). In other words, the existence of a smooth CLF implies smooth stabilizability of the system. Mathematically, a smooth (except possibly at the origin) stabilizer $K(.) : \mathbb{R}^n \to \mathbb{R}^m$ can be constructed provided that a smooth CLF $V : \mathbb{R}^n \to \mathbb{R}$ can be found and say the inequality $\min_u \nabla V(x).f(x,u) \le 0$ can be satisfied by choosing a value of u which depends continuously on the value of x, for each $x \ne 0$ (Bacciotti & Mazzi, 2010). The mentioned Artstein-Sontag problem is expressed as follows:

$$V = IOC(x,f,g) : \begin{array}{c} \min\limits_u \frac{\partial V}{\partial x}\dot{x} < 0 \\ \dot{x} = f(x(t)) + g(x(t))u(t) \\ u(t) = k(x(t)) \\ V > 0,\ \frac{dV}{dt} < 0 \end{array} \tag{10}$$

The control law $k(x(t))$ is obtained by the so-called Sontag formula as:

$$k(x) = \begin{cases} -\frac{a+\sqrt{a^2+b^4}}{b}, & if\ b \ne 0 \\ 0, & if\ b = 0 \end{cases} \tag{11}$$

where $a = \frac{\partial V}{\partial x} f$ and $b = \frac{\partial V}{\partial x} g$. This control law is continuous for all x values within a neighborhood radius from the origin $x = 0$, and it is globally stabilizing if V is a global CLF, and f and g are smooth functions. The goal of this problem is to find a CLF V in such a way that the control law $k(x)$ stabilizes the control system.

Robust IOC: Extending the Artstein-Sontag theorem on the non-linear stabilizability to systems with disturbances leads to the design of a robust IOC. The existence of a robust CLF for a control-affine system declares in the concept of the robust stabilizability. Freeman and Kokotovic (1996) presented an inverse optimal robust stabilization problem for disturbed non-linear systems, as follows:

$$(q,r) = IOCP(x,u) : \begin{array}{c} V(t,x) = \inf\limits_w \min\limits_u \int_0^\infty (q(x) + r(x,u))dt \\ u(t) = -k(x) \\ \dot{x} = f(x(t), u(t), \omega(t)) \\ V > 0,\ \frac{dV}{dt} < 0 \end{array} \tag{12}$$

where the term $\inf\limits_w \min\limits_u \int_0^\infty (q(x) + r(x,u))dt$ looks to maximize the objective function w.r.t. disturbance and minimize it w.r.t. the control variable. The optimal control law is obtained by minimizing the Hamiltonian function $H(x(t), k(x), \omega(t), \frac{\partial V}{\partial x}) = q(x) + r(x,u) + \frac{\partial V}{\partial x}(f(x(t), u(t), \omega(t)))$. By substituting $u(t) = -k(x)$ and forming the HJB equation as $\frac{\partial V}{\partial t} + H(x(t), k(x), \omega(t), \frac{\partial V}{\partial x}) = 0$, the function q, r is obtained as the solution of the IOC problem.

Discrete-time IOC: Ornelas et al. (2010) presented a discrete-time IOC problem for output tracking of a non-linear system based on stabilizing the optimal controller using both CLF and passivity approaches.

$$l(x_k) = IOC(x,u) \quad \begin{array}{c} V(x_k) = \min\limits_u \sum\limits_{k=0}^{\infty} l(x_k) + u^T Ru \\ x_{k+1} = f(x_k) + g(x_k)u_k \\ y_k = h(x_k) + J(x_k)u_k \\ V(x_{k+1}) - V(x_k) \le y_k^T u_k \\ l(x_k) \ge 0 \end{array} \tag{13}$$

where $l(x_k)$ is a positive semidefinite function and the term $V(x_{k+1}) - V(x_k) \le y_k^T u_k$ is the passivity condition. To solve the problem, first, the discrete-time Hamiltonian is formed as $H(x_k, u_k) = l(x_k) + u^T Ru + V(x_{k+1}) - V(x_k)$. Assuming $V(x) = x^T Px$ with a positive definite matrix P and using the Hamiltonian and the passivity condition, the control law is obtained as $u_k = -(I + \frac{1}{2}g^T(x_k))^{-1} g^T(x)Pf(x_k)$, resulting in $l(x_k) = -(f^T(x_k)Pf(x_k) - x_k^T Px)$. Since $l(x_k) \ge 0$, then the use of the passivity condition leads to a passivity-based feedback control law as $u_k = -(I + J(x_k))^{-1} h(x_k)$, where, $h(x_k) = g^T(x_k)P\,f(x_k)$ and $J(x_k) = \frac{1}{2} g^T(x_k)P\,g(x_k)$. Now, given the state and control data in the form of the demonstration trajectories, the matrix P can be achieved.

Residual minimization approach: Keshavarz et al. (2011) proposed a method for imputing or estimating the objective function based on some demonstrated samples. Assuming the optimization problem as follows:

$$\begin{array}{l} \qquad \min f(x,p) \\ s.t.\ g_i(x,p) \le 0,\ i = 1,\dots,m \\ \qquad A(p)x = b(p) \end{array} \tag{14}$$

where $x \in \mathbb{R}^n$ is the variable, $f = \sum_{i=0}^{k} w_i f_i$ and g_i are differentiable and convex for each value of $p \in P$ (the set of allowable parameter values), $A : P \to \mathbb{R}^{q\times n}$, and $b : P \to \mathbb{R}^q$, we let $x \in \mathbb{R}^n$ be optimal for $p \in P$ if it is a solution to the problem expressed as Eq. (14). Here it is not assumed that there is only one solution to Eq. (14) for each p; in other words, there are chances that several x's be optimal for a given p. To solve this problem, the following residuals are obtained:

$$\begin{cases} r_{ineq} = -g_i(x,p) > 0 \\ r_{eq} = A(p)x - b(p) \\ r_{stat}(w,\lambda,\upsilon) = \nabla f(x,p) + \sum\limits_{i=1}^{m} \lambda_i \nabla g_i(x,p) + A(p)^T \upsilon \\ r_{comp}(\lambda) = \lambda_i g_i(x,p) \end{cases} \tag{15}$$

The first two residuals, r_{ineq} and r_{eq}, correspond to primal feasibility and the third residual indicates the stationarity, while the fourth condition is complementary slackness yielded by implementing the necessary and sufficient KKT (Karush-Kohn-Tucker) optimality conditions in a dual problem formulation with the dual parameters λ and υ. Now, given an observation data $D = \{x^k, p^k\},\ k = 1,\dots,N$, the imputed objective problem method involves finding the weights $w \in \mathcal{A}$ in such a way that each decision $x(k)$ is approximately optimal for the associated parameter value $p^{(k)}$. The residual approach to IOC is formulated as follows:

$$(w,\lambda,\upsilon) = IOC(x,p) \quad \begin{array}{c} \min \sum\limits_{k=1}^{N} \left(r_{stat}^{(k)}, r_{comp}^{(k)} \right)_2^2 \\ r_{ineq} = 0 \\ r_{eq} = 0 \\ \lambda_k \ge 0,\ k = 1,\dots,N,\ w \in \mathcal{A} \end{array} \tag{16}$$

This problem is the dual form of the problem Eq. (14) and its goal is to minimize a non-negative convex penalty function $\sum_{k=1}^{N} \|(r_{stat}^{(k)}, r_{comp}^{(k)})\|_2^2$.

Inverse ϵ-optimal control problem: Pauwels et al. (2014) approached to the inverse problem of Lagrangian identification given the system dynamics and optimal trajectories using the HJB sufficient optimality conditions for the direct problem by solving it numerically through polynomial optimization and linear matrix inequalities (LMI). Let $X \subseteq \mathbb{R}^{d_X}$ and $U \subseteq \mathbb{R}^{d_U}$ denote the state and control spaces, respectively, which are supposed to be compact subsets of Euclidean spaces. The dynamic system is assumed to be a continuously differentiable vector field $f \in X \times U$. Also, the terminal state constraints are given by a set of $X_T \subseteq X$ where T is the terminal time. Given a Lagrangian $l(x(t), u(t))$. Accordingly, the forward optimal control problem is formed as below:

$$(u,\ T) = OCP(x,l) : \begin{array}{c} v(t,z) = \inf\limits_{u,T} \int_0^T l(x(t), u(t))dt \\ \dot{x}(t) = f(x(t), u(t)), \\ x(t) \in X,\ u(t) \in U, t \in [0,T], \\ x(0) = z, x(T) \in X_T, \\ T \in [0, T_M] \end{array} \tag{17}$$

Given the observation of the optimal trajectories or demonstration dataset $D = \{(t_i, x_i, u_i)\}_{i\in 1,\dots m} \in ([0,T] \times X \times U)$, by defining a linear operator acting on Lagrangian $l(x(t), u(t))$ and value function

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

v as $\mathcal{L}(l,v) \to l + \frac{\partial v}{\partial t} + \frac{\partial v^T}{\partial t} f$, the IOC problem can be formed in the framework of an inverse ϵ-optimal control problem as follows:

$$(l,v,\epsilon) = IOC(t,x,u):\quad \begin{array}{c} \inf\limits_{l,v,\epsilon} \epsilon + \lambda l(x(t),u(t))_1 \\ \mathcal{L}(l,v)(t,x,u) \geq 0, \forall (t,x,u) \in [0,T] \times X \times U, \\ v(T,x) = 0,\ \ \forall x \in X_T, \\ \frac{1}{n}\sum\limits_{i=1}^{n} \mathcal{L}(l,v)(t_i,x_i,u_i) \leq \epsilon, \\ \mathcal{A}(\mathcal{L}(l,v)) = 1 \end{array} \tag{18}$$

where ϵ is a real value, $\lambda > 0$ is a given regularization parameter, l and v are considered as polynomials, $\|.\|_1$denotes the ℓ_1 norm of a polynomial and $\mathcal{A}$ is a normalization function. The goal of this dual IOC problem is to find the Lagrangian function $l(x(t), u(t))$ and dual parameters v and ϵ given $t, x, u \in D$.

Inverse polynomial optimization: Pauwels et al. (2016) used Lasserre relaxation (Lasserre, 2001) to solve an IOC problem. Consider A as a compact subset of a finite-dimensional Euclidean space $C(A)$ and let $C^1(A)$ be a set of continuous and differentiable function A. Also, let $\mathcal{M}(A)$ be the space of Borel measures on A. Accordingly, the topological dual of $C(A)$ with the duality bracket denoted by $\langle ., .\rangle$, i.e.$\langle \mu, f\rangle = \int A f(x) d\mu(x)$, is the integration of a function $f \in C(A)$ over A w.t.r. a measure $\mu \in M(A)$. Let $M_+(A)$ resp. $C_+(A)$ denote the cone of a non-negative Borel measuring the resp. non-negative continuous functions on A. The support of a measure $\mu \in M+(A)$ is denoted by $spt\,\mu$. An element $\mu \in M_+(A)$ such that $\langle \mu, 1\rangle$ is called a probability measure. Let δx denote the Dirac measure concentrated on x and let $I(e)$ denote the indicator function of an event e, which is equal to 1 if e is true, and 0 otherwise. Using the mentioned assumptions and based on the Lasserre relaxation, the optimal control problem Eq. (17) is translated to the following problem

$$(\mu,\mu_T) = OCP(l,\mu_0):\ s.t. \begin{array}{c} p_0^*(\mu_0) := \inf\limits_{\mu,\mu_T} \mu, l \\ div\ f\mu + \mu_T = \mu_0, \\ \mu, 1 \leq T, \\ \mu \in \mathcal{M}_+(X \times U), \\ \mu_T \in \mathcal{M}_+(X_T) \end{array} \tag{19}$$

where div is divergence operator and μ, μ_0 and μ_T denote the distribution probabilities on the space $[0\,T] \times X \times U$, X_0 and X_T, respectively. Moreover, $\mu(A \times B)$ is called the occupation measure on the set $\times B$. Such a problem can be solved through polynomial optimization (e.g. Lasserre relaxation). Now, the IOC problem in the framework of occupation measures and polynomial optimization can be stated as below:

For $\epsilon > 0$, given measures $\mu \in M_+(C \times U)$ and $\mu_T \in M+(X_T)$ such that $div f \mu + \mu_T \in M_+(X_0)$, where div is the divergence operator, denote by $IOC_\epsilon\ (\mu,\mu_T)$ the set of ϵ-optimal solutions to the IOC problem, namely the set of functions $l \in C(X \times U)$ such that there exists a function $v \in C^1(X)$ satisfying the following:

$$l = IOC_\epsilon(\mu,\mu_T):\ s.t. \begin{array}{c} \mu, l + \nabla v.f \leq \epsilon \\ l + \nabla v.f + \epsilon \in C_+(X \times U), \\ \mu_T, v \geq -\epsilon, \\ -v \in C_+(X_T), \end{array} \tag{20}$$

where ∇ denotes the gradient operator. Using the occupation measure instead of a demonstration set yielded the dual form Eq. (20) for the primal problem Eq. (17). The occupation measure μ is the distribution probability corresponding to the demonstration dataset D. In polynomial optimization, the value function v is approximated by a polynomial function as $v \to v_{\alpha\beta} = \sum_{\alpha+\beta} p_{\alpha\beta} x^\alpha u^\beta$. Using this approximation, the measure problem is relaxed to a moment problem which can be easy to relax to an LMI problem that can be readily solved using the related tools.

Another IOC- related study using the occupation measures was done by Claeys and Sepulchre (2014) who focused o reconstructing the trajectories from the moments of occupation measures inspired by Lasserre relaxation.

5.2. Modern IOC problems

Inverse Reinforcement Learning (IRL): Given a finite space S and a set of k actions $A = \{a_1, a_2, \ldots, a_k\}$, transition probability $\{P_{sa}\} : S \to \mathbb{R}$, policy $\pi : S \to A$ and discount factor $\gamma \in [0, 1)$ and assuming that the model is known and the environment is in the context of Markov decision process (MDP) as $MDP(S,\ A,\ \{P_{sa}\},\ \gamma,\ R)$, a value function is an expectation over cumulative discounted reward for sequences $(s_1, s_2, \ldots)$ passed through when the policy π is implemented from the state s_1 and formed as $V^\pi(s) = E_{s'\sim P_{sa}}[\sum_{t=0}^{\infty} \gamma^t R(s)|\pi] = R(s) + \gamma E_{s'\sim P_{sa}(.)}[\sum_{t=0}^{\infty} V^\pi(s')|\pi]$. A value function represents a long-term criterion for evaluating a learning process, while a reward is a short-time criterion. In addition to the value function, which is merely distributed on states, we need to define a function distributed on both states and actions. In this regard, the state-action Q-function is defined as $Q^\pi(s,a) = R(s) + \gamma E_{s'\sim P_{sa}(.)} V^\pi(s')$, where $s' \sim P_{sa}(.)$ denotes that the next state s' is yielded taking into account the probability distribution $P_{sa}(.)$. The optimal value function and Q-function are declared as $V^*(s) = \sup\limits_{\pi} V^\pi(s)$ and $Q^*(s,a) = \sup\limits_{\pi} Q^\pi(s,a)$, respectively. Given the reward function, the optimization of the policy π by evaluating the value function or Q-function in several episodes is called Reinforcement Learning (RL). Using the transition probability, the value function and Q-function can be rewritten as follows:

$$\begin{array}{l} V^\pi(s) = R(s) + \gamma \sum\limits_{s'} P_{s\pi(s)}\big(s'\big) V^\pi\big(s'\big) \\ Q^\pi(s,a) = R(s) + \gamma \sum\limits_{s'} P_{sa}\big(s'\big) V^\pi\big(s'\big) \end{array} \tag{21}$$

where $\pi(s) \in \arg\max\limits_{a \in A} Q^\pi(s,a)$.

Schaal (1997) introduced an RL from demonstration problem in the context of IOC to learn a balancing task by a robotic arm. In this context, by considering the Q-function $Q(x,\ u)$, value function $V(t,x) = \sum_{k=1}^{\infty} \gamma^{k-1} r(x_{\mathrm{k}},\ u_{\mathrm{k}})$, the policy $\pi(x)$, and the system dynamics $\dot{x} = f(x,u)$, the related LfD problem was formulated as below:

$$(K,\ P, R) = IOC(x,f,\pi)\ s.t. \begin{array}{c} Q(x_k,u_k) = r(x_k,\ u_k) \\ +\gamma \arg\min\limits_{u_{k+1}} (Q(x_{k+1},\ u_{k+1})) \\ x_{k+1} = f(x_k,u_k), \\ \pi(x_k) = -Kx = \arg\min\limits_{u_k} Q(x_k,u_k), \\ r(x_k,\ u_k) = x^T P x + u_k^T R u_k \end{array} \tag{22}$$

The goal of each task was to construct $r(x_k,u_k)$ given a policy $\pi = -Kx$ as the state feedback which minimizes the value-function $V(x_k)$ as the infinite-horizon discounted reward. In addition to the state value function, the process was implemented by considering the state-action function $Q(x_k,\ u_k)$ as a function to be minimized w.t.r. to the control variable u. Assuming the Q-value in a quadratic form as $Q(x,u) = [x^T\ u^T][\begin{smallmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{smallmatrix}][x^T\ u^T]^T,\ H_{21} = H_{12}$, and minimizing $Q(x,\ u)$ w.r.t. the control variable u gives the policy function for direct RL problem as u: $\frac{dQ(x,u)}{du} = 0 \to H_{21}x + H_{22}u = 0 \to u = \pi(x) = -H_{22}^{-1} H_{21} x = Kx$. Reversely, given $K = K^{demo}$ as the control gain related to the expert demonstration, the form of reward function is recovered through several iterations of the learning process as the solution of the LfD problem.

Russell (1998) characterized the IRL for determining the reward function being optimized into three states given 1. mea-

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

surements of an agent's behavior over time in a variety of circumstances, 2. measurements of the sensory inputs to that agent, and 3. a model of the physical environment (including the agent's body). Ng and Russell (2000) proved a theorem claiming that the policy π given by $\pi(s) = a_1$, is optimal if and only if for all actions $a = a_2, \ldots a_k$, the reward R satisfies the inequality $(P_{a1} - P_a)(I - \gamma P_{a1})^{-1} R \geq 0$, where P_a is the transition probability matrix for the action a. This inequality can be proved as follows. For $\forall s \in S$, the transition probability $P_{s\pi(s)}$ is equal to P_{sa}. Since $\pi(s) = a_1$ is assumed as the optimal policy, then it can be stated that $V^{\pi}(s') = V^{\pi}(s)$. It means that the best-predicted value for $V^{\pi}(s')$ is determined by the optimal value function $V^{\pi}(s)$. Therefore, $V^{\pi} = R + \gamma P_{a_1} V^{\pi} \rightarrow (I - \gamma P_{a_1}) V^{\pi} = R \rightarrow V^{\pi} = (I - \gamma P_{a_1})^{-1} R$. On the other hand, the optimal action is obtained as $a_1 = \pi(s, a) \in \arg\max\limits_{a \in A} \sum_{s'} P_{sa}(s') V^{\pi}(s') \forall s \in S$ which gives $\sum_{s'} P_{sa_1}(s') V^{\pi}(s') \geq \sum_{s'} P_{sa}(s') V^{\pi}(s') \ \forall s \in S,\ a \in A$, and subsequently, by simplifying the phrase as $P_{a_1} V^{\pi} \geq P_a V^{\pi} \ \forall s \in S,\ a \in A \backslash a_1$ and substituting $V^{\pi} = (I - \gamma P_{a_1})^{-1} R$, the inequality $(P_{a_1} - P_a)\ (I - \gamma P_{a_1})^{-1} R \geq 0$ is proven for $s \in S$ and $a \in A \backslash a_1$.

Using this theorem and given the above assumptions, the IRL problem is formed to find or reconstruct the unknown reward function $R(s, a) : (S, A) \rightarrow \mathbb{R}$ can be mathematically stated as:

$$R = IRL(S, A, \{P_{sa}\}, \gamma, R) \quad \begin{array}{c} \max \sum\limits_{i=1}^{N} \min\limits_{a \in \{a_1, \ldots, a_k\}} \left\{ (P_{a1}(i) - P_a(i))(I - \gamma P_{a1})^{-1} R \right\} - \lambda R_1 \\ s.t.\ (P_{a1} - P_a)(I - \gamma P_{a1})^{-1} R \geq 0 \ \forall\, a \in A \backslash a_1 \\ |R_i| \leq R_{max}\ ,\ i = 1, \ldots, N \end{array} \tag{23}$$

where, $P_a(i)$ denotes the ith row of the P_a and $-\lambda \|R_1\|$ is a weighted decay-like penalty term with the adjustable penalty coefficient λ for balancing between the goal of having small reinforcements and maximizing the sum of the differences between the quality of the optimal action and the quality of the next best action as $\sum_{s \in S} (Q^{\pi}(s, a_1) - \max\limits_{a \in A \backslash a_1} Q^{\pi}(s, a))$.

Apprenticeship Learning (AL): Abbeel and Ng (2004) proposed a feature-based reward function as $R(s, a) = w^T \phi(s, a)$, where $w \in \mathbb{R}^n$ is a weighting vector and $\phi : (S, A) \rightarrow \mathbb{R}^n$ is a basis function. Using this idea, the value function was rewritten as $V(s_t) = E[\sum_{t=0}^{\infty} \gamma^t w^T \phi(s_t, a_t) | \pi] = w^T \mu(\pi)$, where the new function $\mu(\pi) = E[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | \pi]$ is called feature expectation. The basis function $\phi(s,\ a) \in [0, 1]$ describes whether the constraint for the state s and action a is satisfied or not. For example, considering a constraint set C, it means that the states $s \in S$ are infeasible, then one may say that $\phi(s, a) = 0$ if $s\, or\, s' \in C$ and $\phi(s, a) = 1$ if $s\ \&\ s' \notin C$, where s' is the next state resulted in taking the action a. Following the mentioned introduction, the AL problem is defined as below:

Given an MDP\R, a feature mapping ϕ and the feature expectation μ_E demonstrated by an expert with the policy π^E, the IRL problem is to find the closest policy $\pi \in \Pi$ to that adopted by the expert on the unknown reward function $R = w^T \phi$, where Π is a feasible set of policies. It can be said that, for every $\pi \in \Pi$, the equality $E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | \pi^E] \geq E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | \pi]$ satisfies. Substituting this weighted approximated reward function into this equality leads to the problem of "finding w such that $w^T \mu(\pi^E) \geq w^T \mu(\pi) \forall \pi$ ". Then, the minimization problem can be reduced to an ε-optimal problem as $\|\mu(\pi) - \mu_E\|_2 \leq \varepsilon$. Using this idea, a policy π can be optimized to obtain the closest policy to the expert's policy, and finally, the AL problem is formed to find a reward function such that the expert(teacher) maximally outperforms all previously found controllers as below:

$$(w, \tau, \pi) = IRL(\mu_E) : \ s.t. \ \begin{array}{c} \max\limits_{\tau, w} \ \tau \\ w^T \mu_E \geq w^T \mu^{(j)} + \tau, \ j = 1, 2, \ldots, i - 1 \\ \|w\|_2 \leq 1 \end{array} \tag{24}$$

where τ is a margin denoting the distance to the expert policy (Abbeel & Ng, 2004). The problem expressed by Eq. (24) tries to find the weight $w^{(i)}$ such that $E_{s_0 \sim D}[V^{\pi_E}(s_0)] \geq E_{s_0 \sim D}[V^{\pi^{(i)}}(s_0)] + \tau$. In other words, a reward on which the expert does better, by a "margin" of τ, than any of the i policies found previously.

The max-margin formulation is mutually related to the inverse combinatorial and convex optimization (Burton & Toint, 1992; Ahuja & Orlin, 2001; Heuberger, 2004; Ghobadi et al., 2018) and learning structure prediction models (Taskar et al., 2005). Ratliff et al. (2006a) formed an IL problem as a maximum margin structured prediction problem over a space of policies in the framework of MDP. In a classification task, given a basis function $F : X \times Y \rightarrow \mathbb{R}^2$, a hypothesis $h_w \in H$ is defined as $h_w(x) = \arg\max\limits_{y \in Y} \sum_{i=1}^{n} w_j f_j(x, y) = \arg\max\limits_{y \in Y} w^T F(x, y)$, where w_j is the weighting vector for the jth feature which is extracted to distinguish between some objects. Here the objective is to minimize a loss function $l(x, y, h(x)) = w^T F(x, y)$. The primal and dual problems of Support Vector Machine (SVM, as the main structured prediction problem) can be stated as follows:

$$Primal : \ \begin{array}{c} \min\limits_{w,b} \frac{1}{2} w^T w \\ \left(w x_j + b\right) y_j \geq 1, \ \forall j \end{array} \quad Dual : \ \begin{array}{c} \max\limits_{\alpha} \min\limits_{w,b} L(w, b, \alpha) \\ \alpha \geq 0, \ \forall j \end{array} \tag{25}$$

where $L(w, b, \alpha) = \frac{1}{2} w^T w - \sum_j \alpha_j [(w x_j + b) y_j - 1]$ is a Lagrangian function with multipliers α_j. Given a training dataset $D = \{(X_i,\ A_i,\ p_i,\ f_i,\ y_i, L_i)\}_{i=1}^{n}$, where X_i is the state space, A_i is the action space, p_i is the dynamic model, f_i is the feature function, y_i is an expert demonstration, and L_i is the loss function, feature, and loss functions are linearly decomposed as $f_i = F_i \mu$ and $L_i = F_i l_i$. By considering the case for which both $f_i(\ \cdot\ )$ and $L_i(\ \cdot\ )$ are linear at the state-action frequency μ that they factor over state-action pairs, the maximum margin problem becomes:

$$\begin{array}{l} \min_{w, \xi_i} \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{n} \sum\limits_{i} \beta_i \xi_i \\ s.t.\ \forall\, i\ w^T F_i \mu_i + \xi_i \geq \max_{\mu \in \mathcal{G}} \left(w^T F_i \mu + l_i^T \mu \right) \end{array} \tag{26}$$

where μ_i is the demonstration policy, and $\mu \in \mathcal{G}$ expresses the Bellman-flow constraint for each MDP (*i.e.* $\mu \geq 0$ satisfies $\sum_{x,a} \mu^{x,a} p_i(x'|x, a) + s_i^{x'} = \sum_a \mu^{x',a}$). Also, $p_i(x'|x,\ a)$ is the probability of transition from the current state x to the next state x'upon taking the action a. Computing the dual of the right-hand-side of each constraint gives $w^T F_i \mu(\pi_E^{(i)}) + \xi_i \geq \min\limits_{v \in V_i} s_i^T v$ where $v \in V_i$ are the value-functions that satisfy the Bellman primal constraints $\forall x,\ a\ v^x \geq (w^T F_i + l_i)^{s,a} + \sum_{x'} p_i(x'|x, a) v^{x'}$. Finally, the MMP problem is formed as follows

$$\begin{array}{l} \min_{w, \xi_i, v_i} \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{n} \sum\limits_{i} \beta_i \xi_i \\ s.t.\ \forall\, i\ w^T F_i \mu_i + \xi_i \geq s_i^T v_i \\ \forall\, i,\ x,\ a\ v_i^x \geq \left(w^T F_i + l_i \right)^{s,a} + \sum\limits_{x'} p_i \left(x' | x, a \right) v_i^{x'} \end{array} \tag{27}$$

In this framework, the learner attempts to find a linear mapping of the features to rewards so that, for each problem, the best policy instance over the resulting reward function $\mu^* = argmax_{\mu \in \mathcal{G}} w^T F_i \mu$ is "as close as possible" to the demonstrated policy μ_i, where the closeness is defined by the loss function.

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

Min-Max feature expectation matching. This method was originally proposed by Syed and Schapire (2008) as a min-max IRL problem for learning to play a game:

$$(w,\psi)=IRL(\mu_E):\ s.t.\ \begin{array}{c}\min\limits_{w}\max\limits_{\lambda} w^T G\psi\\ G=w^T(\mu(\pi)-\mu_E)\ \forall\,\pi\\ w\geq 0\\ \|w\|_1=1\end{array} \tag{28}$$

where G is the game matrix and ψ represents the weights which adjust the contributions of different policies $\pi_1,\pi_2,...,\pi_n$ to the mixed policy.

Maximum-entropy feature expectation matching. This method was proposed by Ziebart et al. (2008) who recalled feature matching in a suboptimal expert case to maximize the entropy of distributions over the demonstrated path by satisfying the constraint of feature expectation matching. Based on this approach, the IRL is stated as:

$$(w,\pi)=IRL(\mu_E):\ s.t.\ \begin{array}{c}\max\limits_{P(\xi\in D)}-\sum\limits_{\xi}P(\xi)\,log\,P(\xi)\\ \sum\limits_{\xi}P(\xi)\;\mu(\xi)=\mu(\pi_E)\\ \mu(\xi)=\sum\limits_{t=1}^{\infty}\gamma^t\phi(s_t,a_t)\\ P(\xi)=\frac{1}{Z(w)}exp\left(w^T\mu(\xi)\right)\\ Z(w)=\sum\limits_{\xi'\in D}exp\left(w^T\mu\left(\xi'\right)\right)\\ \sum\limits_{\xi}P(\xi)=1\end{array}\quad\forall\,\pi \tag{29}$$

where $P(\xi)$ is the distribution probability of the trajectories' demonstrated dataset $D=\{\xi_1,\xi_2,\ldots,\xi_n\}$, with individual trajectories $\xi_j=\{s_0,\ a_0,s_1,a_1,\ldots,s_{k_j}\}_j^{k_j}\subseteq D$. It is assumed that the expert stochastically chooses between paths ξ where each path's logarithmic probability is given by its expected sum of rewards.

IOC with locally optimal examples: Levine and Koltun (2012) introduced a probabilistic IOC algorithm using a local approximation of reward for continuous, high-dimensional domains. The problem was stated as below:

$$r=IRL(x,u):\ s.t.\ \begin{array}{c}\max\limits_{u}\ \mathcal{L}=\log p(u|x_0)\\ x_{t+1}=f(x_t,u_t)\\ u=\arg\max\limits_{u}\sum\limits_{t}r(x_t,u_t)\\ p(u|x_0)=\frac{1}{Z}exp\left(\sum\limits_{t}r(x_t,u_t)\right)\\ \sum\limits_{u}P(u|x_0)=1\end{array} \tag{30}$$

Where, Z is the partition function. To solve the problem for obtaining the reward function likelihood, the probability $P(u|x_0)$ was approximated using a second-order Taylor expansion of the reward function $r(x_t,\ u_t)$ around u as $r(\tilde{u})\approx r(u)+(\tilde{u}-u)^T\frac{\partial r}{\partial u}+\frac{1}{2}(\tilde{u}-u)^T\frac{\partial^2 r}{\partial u^2}(\tilde{u}-u)$. Denoting the gradient $\frac{\partial r}{\partial u}$ as g and the Hessian $\frac{\partial^2 r}{\partial u^2}$ as H, the mentioned expansion yielded to obtain $P(u|x_0)\approx\exp(\frac{1}{2}g^TH^{-1}g|-H|^{\frac{1}{2}}(2\pi)^{-\frac{du}{2}})$. Then the likelihood function was obtained as $\mathcal{L}=\frac{1}{2}g^TH^{-1}g+\frac{1}{2}\log|-H|-\frac{d_u}{2}\log 2\pi$ and then used to find a solution to the problem.

Mombaur et al. (2010) proposed a bi-level method to find a reward function for a deterministic system $x_{t+1}=f(x_t,u_t)$ learning a demonstrated expert trajectory by minimizing a least-square loss function:

$$w=IRL(\pi_E)\ s.t.\ \begin{array}{rl}Upper\ level: & \min\limits_{w}\sum\limits_{t=0}^{T}v_t^{\pi_E}-v_t^{\pi_w}{}_2\\ Lower\ level: & \begin{cases}v_t^{\pi_w}=\min\limits_{z,u,T}\sum\limits_{t=0}^{T}w^T\phi(z_t,u_t)\\ z_{t+1}=f(z_t,u_t)\\ z_0=z_0^E,\ \ z_T=z_T^E\end{cases}\end{array} \tag{31}$$

where z and u were state and control variables, respectively. $v_t^{\pi_E}$ was the expert demonstration, and $v_t^{\pi_w}$ denoted the trajectory learned by the learning agent with the matrix $v=(z_t^T,\ u^T)$ that was approximated by a linear function of the weighting matrix w and the feature vector ϕ. A bi-level method was proposed to solve this problem including an upper level that handled the iteration over w such that the fit between the demonstration trajectory and optimal control problem solution was improved. Each upper-level iteration included one call to the lower level where a forward optimal control problem was solved for the current set of w_i. The optimal solution of this problem was then applied to the upper level such that the least-squares fit between demonstrated trajectory and computations could be evaluated for this iteration, and finally the weighting matrix w and a parameterized policy class were obtained as $\pi_w=\arg\min_w\sum_{t=0}^{T}\|v_t^{\pi_E}-v_t^{\pi_w}\|_2$.

Dvijotham and Todorov (2010) presented an algorithm for IOC within the framework of linearly solvable MDPs to recover the policy, value function, and cost function. The problem used in this method can be stated as follows:

$$(w,\phi)=IRL(x,)\quad\begin{array}{c}min_w\ L(w,\theta)=\sum_n(w^T\phi(x_n';\theta)+logG(x_n))\\ \phi_i(x;\theta)=\frac{exp(\theta_i^Ts(x))}{\sum_j exp(\theta_i^Ts(x))}\\ s(x)=[1;x_k;x_kx_l]\forall k\leq l\\ G(x)=\int p(x'|x)e^{-w^T\phi(x;,\theta)}dx'\\ \pi(x'|x)=\frac{p(x'|x)e^{-w^T\phi(x;,\theta)}}{\int p(x'|x)e-w^T\phi(x;,\theta)}dx'\end{array} \tag{32}$$

where x and x' were current and next sates, respectively, $p(x'|x)$ and $\pi(x'|x)$ were state transition probability and policy corresponding to the MDP, respectively, and $\phi(x_n';\theta)$ and $G(x)$ were feature vector and normalizing function, respectively. Herein w denotes a vector of linear weights while θ is a vector of parameters that affect the shape and location of the bases ϕ_i.

Generative adversarial IRL: Ho and Ermon (2016) proposed a new approach called Generative Adversarial Imitation Learning (GAIL) to recover a policy that matched the expert demonstrations by optimizing a cost learned through MaxEnt IRL (Ziebart et al., 2008; Ziebart, 2010) to find a cost function from a family of functions $c\in C$ such that a given expert policy π_E is uniquely optimal w.r.t that cost function by using the optimization problem.

$$\max_{c\in C}\left(\min_{\pi\in\Pi}-H(\pi)\ +E_\pi[c(s,a)]\right)-E_{\pi_E}[c(s,a)] \tag{33}$$

where $H(\pi)\triangleq E_\pi[-log\pi(a|s)]$ is the γ-discounted causal entropy of the policy π. Maximum causal entropy IRL find a cost function $c\in C$ in such a way that minimizes the cost to the expert policy and maximize the cost to other policies, and by using it, the expert policy can be found in a framework of RL as $RL(c)=arg\min\limits_{\pi\in\Pi}-H(\pi)\ +E_\pi[c(s,a)]$ by forming a cost function into high-entropy policies minimizing the expected cumulative cost. In case of search for an IL algorithm in a large environment with a large set of cost functions, to avoid overfitting the RL, Ho and Ermon (2016) used a feature-based cost function as a convex cost function regularizer $\psi:\mathbb{R}^{S\times A}\rightarrow\mathbb{R}$ and defined an IRL primitive procedure with the cost regularized by ψ as follows:

$$\begin{aligned}IRL_\psi(\pi_E)=\arg\max_{c\in\mathbb{R}^{S\times A}}-\psi(c)+\Big(&\min_{\pi\in\Pi}-H(\pi)\\ &+E_\pi[c(s,a)]\Big)-E_{\pi_E}[c(s,a)]\end{aligned} \tag{34}$$

If $\hat{c}\in IRL\psi(\pi_E)$ is a cost function learned by the mentioned procedure under the policy π_E and regularization function ψ, by defining the occupancy measure $\rho_\pi:S\ \times A\ \rightarrow\mathbb{R}$ as $\rho_\pi(s,a)=\pi(a|s)\sum_{t=0}^{\infty}=\gamma^tP(s_t=s|\pi)$ for a policy $\pi\ \in\ \Pi$ and rewriting $E_\pi[c(s,a)]\ =\sum_{s,a}\rho_\pi(s,a)c(s,a)$, the characterized reinforcement

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

learning $RL(\hat{c})$ is obtained as follows:

$$RL^{\circ}IRL\psi(\pi_E) = arg\min_{\pi\in\Pi} -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E}) \tag{35}$$

This formulation seeks a policy whose occupancy measure is as close as possible to the expert's policy (i.e. the induced optimal policy) as the primal optimum that is yielded by running RL after IRL. The AL generative model is an alternative to the typical IRL that tries to induce a policy that matches the expert's occupancy measures. Also, the dual form of the optimization problem Eq. (33) is as follows:

$$\begin{aligned} &\min_{\rho\in D} -\bar{H}(\rho) \\ &s.t.\ \rho(s,a) = \rho E(s,a)\ \forall\, s\in S,\ a\in A \end{aligned} \tag{36}$$

So, it can be said that IRL is a dual of an occupancy measure matching with the problem. Since an AL needs to craft features very carefully and also, the procedure $RLIRL\psi(\pi_E)$ encodes the expert's behavior into C, and decoding it back might not be possible if C is too simple, Ho and Ermon (2016) used a more expressive class of cost function as:

$$\psi_{GA}(c) \triangleq \begin{cases} E_{\pi_E}[g(c(s,a))] & if\ \ c < 0 \\ +\infty & otherwise \end{cases} \ where,$$

$$g(x) = \begin{cases} -x - log(1-e^x) & if\ x < 0 \\ +\infty & otherwise \end{cases} \tag{37}$$

In the framework of generative adversarial networks (Goodfellow et al., 2014), similar to IRL, GANs learn an objective for generative modeling. It works by simultaneously training two models including a generator G and a discriminator D. The discriminator classifies its inputs as either the output of the generator or data samples with distribution $p(x)$. Accordingly, the demonstrated trajectory ξ, the policy $\pi \sim q(\xi)$, and the reward R are translated into the GANs model as the sample x, the generator G, and the discriminator D, respectively. For a fixed generator with a [typically unknown] density $q(\xi)$, the optimal discriminator is defined as $D^*(\tau) = \frac{p(\xi)}{p(\xi)+q(\xi)}$, where $p(\xi)$ is the actual distribution of the data. To estimate the discriminator D, a new discriminator D_θ with the parameter θ is defined as $D_\theta(\tau) = \frac{\hat{p}_\theta(\xi)}{\hat{p}(\xi)+q(\xi)}$. Using this introduction, the GAIL is formulating as follows:

$$\psi^*_{GA}(\rho_\pi - \rho_{\pi_E}) = \sup_{D\in(0,1)^{S\times A}} E_\pi[log(D(s,a))] + E_{\pi_E}[log(1-D(s,a))] \tag{38}$$

where ψ^*_{GA} is the optimal negative logarithmic loss of the binary classification problem of distinguishing between state-action pairs of π and π_E.

Ho and Ermon (2016) and Magni and Sepulchre (1997) further used the typical unconstrained form of the discriminator rather than the generator's density. So, the cost function remained implicit within the discriminator and could not be recovered. Hence, in GAIL, the discriminator is discarded and the policy is the final result (Finn et al., 2016a). Following this work, To make the connection to MaxEnt IRL, Finn et al. (2016b) replaced the estimated data density with the Boltzmann distribution as $D_\theta(\tau) = \frac{\frac{1}{z}exp(-c(\tau))}{\frac{1}{z}exp(-c(\tau))+q(\xi)}$. Henderson et al. (2018) used the GAIL framework (Ho & Ermon, 2016) and formulated a method for learning joint reward-policy options with adversarial methods in IRL called OptionGAN. This method could implicitly learn divisions in the demonstration-state space and accordingly learn policy and reward options in one shot. Following a similar approach, Torabi et al. (2018) developed a model to learn a particular task by observing an expert performing that task without the knowledge of the specific actions taken in the framework of BC as a supervised learning scheme, FEM as an AL method, and GAIL. Sun and Ma (2019) proposed a new algorithm called Action-Guided Adversarial Imitation Learning (AGAIL) that learned a policy from incomplete demonstrations by separating the demonstrations into state and action trajectories and training the policy with the state trajectories while using the action trajectories as auxiliary information to guide the training.

6. Challenges

The IRL suffers from some challenges and undesired characteristics of real worlds appearing as curses. Some of the challenges are indicated in the following.

- Ill-posedness of IRL problem

Jacques Hadamard (1865–1963) believed that a well-posed inverse problem is the one that (1) its solution exists, (2) the solution is unique, and (3) the solution depends continuously on the data. In contrast, a problem is ill-posed if it satisfies none of the three conditions for well-posedness. Since the inverse problems are often ill-posed, then they are typically harder to solve numerically, as compared to the forward problems. Therefore, it means that they need to be modified as a new form called the regularization of the problem. The main challenging issue is that the regularization does not guarantee the global solution of a problem. For example, the proposed algorithm by Abbeel and Ng (2004) for addressing the AL cannot guarantee the true recovery of the expert's reward function.

- Non-convexity

An optimization problem is said to be convex if the related objective function and constraints (*e.g.* system dynamics and state-control constraints) are convex functions. Conversely, an optimization problem that violates either one or both of these conditions, i.e. one that has a non-convex objective and/or a non-convex constraint set, is called a non-convex optimization problem (Jain & Kar, 2017). In many natural problems, the objective function is non-convex. Although the non-convex constraints and the objective function provide a more accurate model of the present problem, the convexity itself is a challenging issue, as the non-convexity makes it more difficult to solve the optimization problem. A wide range of problems in the areas of machine learning and data-driven control engineerings such as robotic operations, image processing, driving unmanned vehicles and data mining come with large dimensions and enormous amounts of data. In the face of such systems, it is necessary to impose structural modifications onto the conventional learning patterns. Applying such modifications is not only useful for the regularization of the learning problem but also advantageous for preventing its ill-posedness. Regardless of these modifications (e.g. approximation), the problem remains in the non-convex form.

- Data availability

In IL or LfD, it is typically assumed that the demonstration data is available, while there are cases where the data is incomplete or inaccurate due to uncertainty (Ross et al., 2011; Chernova and Veloso, 2009; Argall, Browning, & Veloso, 2011). Such problems may incrementally query the expert inappropriate regions of the state space to improve the learned policy or to reduce the uncertainty by assuming that (1) the expert exhibits optimal behavior, (2) the expert demonstrations are abundant, and (3) the expert stays with the learning agent throughout the training. In practice, however, these assumptions significantly limit the applicability of the LfD (Kim, Farahmand, Pineau, & Precup, 2013). To query new data to decrease the uncertainty, one needs forward learning methods such as the RL with a combination of expert and interaction data (*i.e.*, mixing LfD and RL) to address the

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

challenging real-world policy learning problems under realistic assumptions. In such cases, where new data is constantly generated and become available, such assumptions cannot be met. Besides, even data arriving at unrealistic time intervals can be a challenge (L'heureux, Grolinger, Elyamany, & Capretz, 2017).

- Non-linearity, complexity and dimensionality curses

In many applications, learning problems are faced with curses of non-linearity, complexity, and dimensionality. For example, in the robotics, the large number of degrees of freedom (DOF) and unknown and uncertain non-linear dynamics are some major challenges for IOC as this control method is based on the solution of the optimal control problem, which is difficult to handle for high-level problems which are frequently solved with no guarantee of optimality (Byravan, Montfort, Ziebart, Boots, & Fox, 2014) even though discretizing such non-linear systems with high-level parameters and big data leads to increased dimensionality and complexity of the related problems. In most cases, especially in classic applications, the system dynamics and cost function are assumed to be linear and quadratic, respectively, while these assumptions may rarely render useful for systems such as robots operating in such a dynamic and complicated environment, in practice. Because of the related environment, workspace, obstacles, and constraints are often multidimensional and non-linear. Non-convexity and discontinuities in the cost function and constraints make it difficult and sometimes impossible to solve the optimal control problem. Challenges such as dimensionality, high DOF, non-linearity of dynamics and cost functions, and non-convexity of constraints tend to make the LfD problem relatively unreliable.

- Feature selection

Feature selection aims to select the most relevant features to reduce the dimensionality of a learning problem and hence save the learning time. With problems of higher dimensions, it is challenging due to spurious correlations and incidental endogeneity (correlation of an explanatory variable with the error term) (Fan, Han, & Liu, 2014; L'heureux et al., 2017). In IRL, such methods as projection, max-margin, MMP, MWAL, LEARCH, and MLIRL are highly sensitive to the feature selection (Arora & Doshi, 2018). Many information serves as feature-based data for a learning problem, while some of them may render not useful, redundant, or not linearly independent of the other variables. To find the relevant features, one needs to consider the low-rank problem, though it leads to ill-posedness of the problem and increased complexity.

- Generalizability

Within the framework of LfD, a cost function is recovered using some trajectories data. In reality, the recovered cost function of a policy is just valid for the demonstrated trajectories while the common practice is to generalize the results to the entire state-action space, which is an important issue in the learning problems. The challenge is to generalize the results correctly to the non-observed space using the data that often covers only a fraction of the complete space (Arora & Doshi, 2018).

7. Conclusion

In this review, the IOC/IRL approaches to the LfD problem were addressed. A historical review of such problems was presented and the categorization of the related methods was provided. The challenging issues associated with the IOC/IRL were further discussed. This review is useful for researchers who re to further investigate the IOC and IRL problems. The primary significance of this review is the proper clarification of the relationship between IOC and IRL problems. This article can be a good reference for future studies to bring these two approaches closer together.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

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

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. *Proceedings of the twenty-first international conference on machine learning* p. 1. ACM.

Aghasadeghi, N. (2015). *Inverse optimal control for differentially flat systems with application to lower-limb prosthetic devices*. Doctoral dissertation, University of Illinois at Urbana-Champaign.

Aghasadeghi, N., & Bretl, T. (2011). Maximum entropy inverse reinforcement learning in continuous state spaces with path integrals. In *2011 IEEE/RSJ international conference on intelligent robots and systems* (pp. 1561–1566). IEEE.

Aghasadeghi, N., Long, A., & Bretl, T. (2012). Inverse optimal control for a hybrid dynamical system with impacts. In *2012 IEEE international conference on robotics and automation* (pp. 4962–4967). IEEE.

Ahuja, R. K., & Orlin, J. B. (2001). Inverse optimization. *Operations Research, 49*(5), 771–783.

Akhiezer, N. I. (1962). *The calculus of variations*. Blaisdell, New York: Google Scholar IM Gel'fand and SV Fomin, The Calculus of Variations.

Almobaied, M., Eksin, I., & Guzelkaya, M. (2015). A new inverse optimal control method for discrete-time systems. In *2015 12th international conference on informatics in control, automation and robotics (ICINCO): 1* (pp. 275–280). IEEE.

Almobaied, M., Eksin, I., & Guzelkaya, M. (2018). Inverse optimal controller based on extended Kalman filter for discrete-time nonlinear systems. *Optimal Control Applications and Methods, 39*(1), 19–34.

Al-Tamimi, A., Lewis, F. L., & Abu-Khalaf, M. (2008). Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 38*(4), 943–949.

Anderson, B. D. (1966). *THE inverse problem of optimal control (No. TR-6560-3)*. Stanford Univ Calif Stanford Electronics Labs.

Anderson, B. D., & Moore, J. B. (2007). *Optimal control: Linear quadratic methods*. Courier Corporation.

Anderson, B. D. O., & Moore, J. B. (1989). *Optimal Control: Linear Quadratic Methods*. Upper Saddle River: Prentice-Hall, Inc.

Arbel, A., & Gupta, N. K. (1981). Robust colocated control for large flexible space structures. *Journal of Guidance and Control, 4*(5), 480–486.

Arechavaleta, G., Laumond, J. P., Hicheur, H., & Berthoz, A. (2008). An optimality principle governing human walking. *IEEE Transactions on Robotics, 24*(1), 5–14.

Argall, B. D., Browning, B., & Veloso, M. M. (2011). Teacher feedback to scaffold and refine demonstrated motion primitives on a mobile robot. *Robotics and Autonomous Systems, 59*(3–4), 243–255.

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

Arora, S., & Doshi, P. (2018). A survey of inverse reinforcement learning: Challenges, methods and progress. arXiv preprint arXiv:1806.06877.

Artstein, Z. (1983). Stabilization with relaxed controls. *Nonlinear Anal. TMA, 7*, 1163–1173.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine, 34*(6), 26–38.

Audiffren, J., Valko, M., Lazaric, A., & Ghavamzadeh, M. (2015). Maximum entropy semi-supervised inverse reinforcement learning. *Twenty-fourth international joint conference on artificial intelligence*.

Babes, M., Marivate, V., Subramanian, K., & Littman, M. L. (2011). Apprenticeship learning about multiple intentions. In *Proceedings of the 28th International conference on machine learning (ICML-11)* (pp. 897–904).

Bacciotti, A., & Mazzi, L. (2010). From Artstein-Sontag theorem to the min-projection strategy. *Transactions of the Institute of Measurement and Control, 32*(6), 571–581.

Bagnell, J. A., Chestnutt, J., Bradley, D. M., & Ratliff, N. D. (2007). Boosting structured prediction for imitation learning. In *Advances in neural information processing systems* (pp. 1153–1160).

Bain, M., & Sommut, C. (1999). A framework for behavioural claning. *Machine Intelligence, 15*(15), 103.

Bakker, P., & Kuniyoshi, Y. (1996). Robot see, robot do: An overview of robot imitation. In *AISB96 workshop on learning in robots and animals* (pp. 3–11).

Bandera, J. P., Rodriguez, J. A., Molina-Tanco, L., & Bandera, A. (2012). A survey of vision-based architectures for robot learning by imitation. *International Journal of Humanoid Robotics, 9*(01), Article 1250006.

Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, 679–684.

Bellman, R. (1970). Dynamic programming and inverse optimal problems in mathematical economics. *Journal of Mathematical Analysis and Applications, 29*(2), 424–428.

Bellman, R., & Dreyfus, S. (1959). Functional approximations and dynamic programming. In *Mathematical tables and other aids to computation* (pp. 247–251).

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

Bellman, R., & Kalaba, R. (1963). *An inverse problem in dynamic programming and automatic control*. CA: Rand Corp Santa Monica (No. RAND-RM-3592-PR).

Bertsekas, D. P., & Tsitsiklis, J. N. (1995). Neuro-dynamic programming: An overview. In *Proceedings of 1995 34th IEEE conference on decision and control: 1* (pp. 560–564). IEEE.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Billard, A., & Grollman, D. (2013). Robot learning by demonstration. *Scholarpedia, 8*(12), 3824.

Billard, A., & Matarić, M. J. (2000). A biologically inspired robotic model for learning by imitation. In *Proceedings of the fourth international conference on autonomous agents* (pp. 373–380). ACM.

Billard, A., & Matarić, M. J. (2001). Learning human arm movements by imitation:: Evaluation of a biologically inspired connectionist architecture. *Robotics and Autonomous Systems, 37*(2–3), 145–160.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blajer, W., Goszczyński, J. A., & Krawczyk, M. (2002). The inverse simulation study of aircraft flight path reconstruction. *Transport, 17*(3), 103–107.

Bliss, G.A. (1946). Lectures on the Calculus of Variations.

Bogdanovic, M., Markovikj, D., Denil, M., & De Freitas, N. (2015). Deep apprenticeship learning for playing video games. *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.

Bolza, O. (1909). *Vorlesungen über variationsrechnung*. Leipzig: Koehler u. Amelang.

Boularias, A., Kober, J., & Peters, J. (2011). Relative entropy inverse reinforcement learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 182–189).

Boyd, S., El Ghaoui, L., Feron, E., & Balakrishnan, V. (1994). *Linear matrix inequalities in system and control theory*: Vol. 15. Siam.

Brown, D.S., Goo, W., Nagarajan, P., & Niekum, S. (2019). Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. arXiv preprint arXiv:1904.06387.

Bryson, A. E. (1996). Optimal control-1950 to 1985. *IEEE Control Systems Magazine, 16*(3), 26–33.

Burger, M., & Osher, S. J. (2005). A survey on level set methods for inverse problems and optimal design. *European Journal of Applied Mathematics, 16*(2), 263–301.

Burton, D., & Toint, P. L. (1992). On an instance of the inverse shortest paths problem. *Mathematical Programming, 53*(1–3), 45–61.

Busby, H. R., & Trujillo, D. M. (1997). Optimal regularization of an inverse dynamics problem. *Computers & Structures, 63*(2), 243–248.

Byravan, A., Montfort, M., Ziebart, B., Boots, B., & Fox, D. (2014). Layered hybrid inverse optimal control for learning robot manipulation from demonstration. *NIPS workshop on autonomous learning robots*.

Byrne, R. W., & Russon, A. E. (1998). Learning by imitation: A hierarchical approach. *Behavioral and Brain Sciences, 21*(5), 667–684.

Calinon, S., & Billard, A. G. (2007). What is the teacher's role in robot programming by demonstration?: Toward benchmarks for improved learning. *Interaction Studies, 8*(3), 441–464.

Casti, J. (1980). On the general inverse problem of optimal control theory. *Journal of Optimization Theory and Applications, 32*(4), 491–497.

Casti, J.L. (1974). A Note on the General Inverse Problem of Optimal Control Theory.

Chang, F. R. (1988). The inverse optimal problem: A dynamic programming approach. *Econometrica, 56*(1), 147 (1986-1998).

Chen, C. C., & Shaw, L. (1982). On receding horizon feedback control. *Automatica, 18*(3), 349–352.

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

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

Choi, J., & Kim, K. E. (2013). Bayesian nonparametric feature construction for inverse reinforcement learning. *Twenty-third international joint conference on artificial intelligence*.

Choi, S., Lee, K., & Oh, S. (2019). Robust learning from demonstrations with mixed qualities using leveraged gaussian processes. *IEEE Transactions on Robotics, 35*(3), 564–576.

Claeys, M., & Sepulchre, R. (2014). Reconstructing trajectories from the moments of occupation measures. In *53rd IEEE conference on decision and control* (pp. 6677–6682). IEEE.

Clever, D., & Mombaur, K. D. (2016). An inverse optimal control approach for the transfer of human walking motions in constrained environment to humanoid robots. *Robotics: Science and systems*.

Curtis III, J. W. (2002). *A generalization of Sontag's formula for high-performance CLF-based control*. Department of Electrical and Computer Engineering, Brigham Young University.

Daftry, S., Bagnell, J. A., & Hebert, M. (2016). Learning transferable policies for monocular reactive mav control. In *International symposium on experimental robotics* (pp. 3–11). Springer.

De Farias, D. P., & Van Roy, B. (2006). A cost-shaping linear program for average-cost approximate dynamic programming with performance guarantees. *Mathematics of Operations Research, 31*(3), 597–620.

Delson, N., & West, H. (1996). Robot programming by human demonstration: Adaptation and inconsistency in constrained motion, *1*, 30–36.

Deng, H., & Krstić, M. (1997). Stochastic nonlinear stabilization—II: Inverse optimality. *Systems & Control Letters, 32*(3), 151–159.

Deniša, M., Gams, A., Ude, A., & Petrič, T. (2015). Learning compliant movement primitives through demonstration and statistical generalization. *IEEE/ASME Transactions on Mechatronics, 21*(5), 2581–2594.

Doerr, A., Ratliff, N. D., Bohg, J., Toussaint, M., & Schaal, S. (2015). Direct loss minimization inverse optimal control. *Robotics: Science and Systems*.

Duan, Y., Andrychowicz, M., Stadie, B., Ho, O. J., Schneider, J., & Sutskever, I. (2017). One-shot imitation learning. In *Advances in neural information processing systems* (pp. 1087–1098).

Dulikravich, G. S. (1988). Inverse design and active control concepts in strong unsteady heat conduction. *Applied Mechanics Reviews, 41*(6), 270–277.

Dvijotham, K., & Todorov, E. (2010). Inverse optimal control with linearly-solvable MDPs. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 335–342).

Englert, P., Paraschos, A., Deisenroth, M. P., & Peters, J. (2013). Probabilistic model-based imitation learning. *Adaptive Behavior, 21*(5), 388–403.

Englert, P., Vien, N. A., & Toussaint, M. (2017). Inverse KKT: Learning cost functions of manipulation tasks from demonstrations. *The International Journal of Robotics Research, 36*(13–14), 1474–1488.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review, 1*(2), 293–314.

Fausz, J. L., Chellaboina, V. S., & Haddad, W. M. (2000). Inverse optimal adaptive control for non-linear uncertain systems with exogenous disturbances. *International Journal of Adaptive Control and Signal Processing, 14*(1), 1–38.

Finn, C., Christiano, P., Abbeel, P., & Levine, S. (2016). A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. arXiv preprint arXiv:1611.03852.2016b.

Finn, C., Levine, S., & Abbeel, P. (2016b). Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning* (pp. 49–58). 2016a.

Finn, C., Yu, T., Zhang, T., Abbeel, P., & Levine, S. (2017). One-shot visual imitation learning via meta-learning. arXiv preprint arXiv:1709.04905.

Freeman, R., & Kokotovic, P. V. (2008). *Robust nonlinear control design: State-space and lyapunov techniques*. Springer Science & Business Media.

Freeman, R. A., & Kokotovic, P. V. (1996). Inverse optimality in robust stabilization. *SIAM Journal on Control and Optimization, 34*(4), 1365–1391.

Freeman, R. A., & Primbs, J. A. (1996). Control Lyapunov functions: New ideas from an old source. In *Proceedings of 35th IEEE conference on decision and control: 4* (pp. 3926–3931). IEEE.

Fujii, T. (1987). A new approach to the LQ design from the viewpoint of the inverse regulator problem. *IEEE Transactions on Automatic Control, 32*(11), 995–1004.

Fujii, T., & Khargonekar, P. P. (1988). Inverse problems in H/sub infinity/control theory and linear-quadratic differential games. In *Proceedings of the 27th IEEE conference on decision and control* (pp. 26–31). IEEE.

Fujii, T., & Narazaki, M. (1984). A complete optimality condition in the inverse problem of optimal control. *SIAM journal on Control and Optimization, 22*(2), 327–341.

Gao, Y., Peters, J., Tsourdos, A., Zhifei, S., & Joo, E. M. (2012). *International Journal of Intelligent Computing and Cybernetics, 5*(3), 293–311.

Gaurav, S., & Ziebart, B. (2019). Discriminatively learning inverse optimal control models for predicting human intentions. *Proceedings of the 18th international conference on autonomous agents and multiagent systems* (pp. 1368–1376). International Foundation for Autonomous Agents and Multiagent Systems.

Ghalamzan, E., Amir, M., Paxton, C., Hager, G. D., & Bascetta, L. (2015). An incremental approach to learning generalizable robot tasks from human demonstration. *2015 IEEE international conference on robotics and automation (ICRA)*.

Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2015). Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning, 8*(5–6), 359–483.

Ghobadi, K., Lee, T., Mahmoudzadeh, H., & Terekhov, D. (2018). Robust inverse optimization. *Operations Research Letters, 46*(3), 339–344.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In *Advances in neural information processing systems* (pp. 3909–3917).

Harder, F., & Wachsmuth, G. (2019). Optimality conditions for a class of inverse optimal control problems with partial differential equations. *Optimization, 68*(2–3), 615–643.

Hayes, G. M., & Demiris, J. (1994). *A robot controller using learning by imitation*. University of Edinburgh, Department of Artificial Intelligence (pp. pp-198).

Henderson, P., Chang, W. D., Bacon, P. L., Meger, D., Pineau, J., & Precup, D. (2018). Optiongan: Learning joint reward-policy options using generative adversarial inverse reinforcement learning. *Thirty-Second AAAI conference on artificial intelligence*.

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., et al. (2018). Rainbow: Combining improvements in deep reinforcement learning. *Thirty-second AAAI conference on artificial intelligence*.

Heuberger, C. (2004). Inverse combinatorial optimization: A survey on problems, methods, and results. *Journal of Combinatorial Optimization, 8*(3), 329–361.

Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. In *Advances in neural information processing systems* (pp. 4565–4573).

Huang, B., Ma, X., & Vaidya, U. (2019). Data-driven nonlinear stabilization using koopman operator. arXiv preprint arXiv:1901.07678.

Hussein, A., Gaber, M. M., Elyan, E., & Jayne, C. (2017). Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR), 50*(2), 21.

Ijspeert, A. J., Nakanishi, J., & Schaal, S. (2003). Learning attractor landscapes for learning motor primitives. In *Advances in neural information processing systems* (pp. 1547–1554).

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

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

Iwamoto, S. (1976). Inverse dynamic programming. Memoirs of the Faculty of Science, Kyushu University. *Series A, Mathematics, 30*(1), 25–42.

Iyengar, G., & Kang, W. (2005). Inverse conic programming with applications. *Operations Research Letters, 33*(3), 319–330.

Jain, P., & Kar, P. (2017). Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning, 10*(3–4), 142–336.

Jameson, A., & Kreindler, E. (1973). Inverse problem of linear optimal control. *SIAM Journal on Control, 11*(1), 1–19.

Jankovic, M., & Kolmanovsky, I. (2000). Constructive Lyapunov control design for turbocharged diesel engines. *IEEE Transactions on Control Systems Technology, 8*(2), 288–299.

Jin, M., Damianou, A., Abbeel, P., & Spanos, C. (2015). Inverse reinforcement learning via deep gaussian process. arXiv preprint arXiv:1512.08065.

Johnson, M., Aghasadeghi, N., & Bretl, T. (2013). Inverse optimal control for deterministic continuous-time nonlinear systems. In *52nd IEEE conference on decision and control* (pp. 2906–2913). IEEE.

Kalakrishnan, M., Pastor, P., Righetti, L., & Schaal, S. (2013, May). Learning objective functions for manipulation. In *2013 IEEE international conference on robotics and automation* (pp. 1331–1336). IEEE.

Kalman, R. E. (1964). When is a linear control system optimal? *Journal of Basic Engineering, 86*(1), 51–60.

Kawasaki, N., & Shimemura, E. (1983). Determining quadratic weighting matrices to locate poles in a specified region. *Automatica, 19*(5), 557–560.

Kawato, M., Gandolfo, F., Gomi, H., & Wada, Y. (1994). Teaching by showing in kendama based on optimization principle. In *International conference on artificial neural networks* (pp. 601–606). Springer.

Keshavarz, A., Wang, Y., & Boyd, S. (2011). Imputing a convex objective function. In *2011 IEEE international symposium on intelligent control* (pp. 613–619). IEEE.

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

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

Kim, B., Farahmand, A. M., Pineau, J., & Precup, D. (2013). Learning from limited demonstrations. In *Advances in neural information processing systems* (pp. 2859–2867).

Klein, E., Geist, M., Piot, B., & Pietquin, O. (2012). Inverse reinforcement learning through structured classification. In *Advances in neural information processing systems* (pp. 1007–1015).

Kogan, M. M. (1997). A local approach to solving the inverse minimax control problem for discrete-time systems. *International Journal of Control, 68*(6), 1437–1448.

Kolter, J. Z., Abbeel, P., & Ng, A. Y. (2008). Hierarchical apprenticeship learning with application to quadruped locomotion. In *Advances in neural information processing systems* (pp. 769–776).

Krejci, P., & Kuhnen, K. (2001). Inverse control of systems with hysteresis and creep. *IEE Proceedings-Control Theory and Applications, 148*(3), 185–192.

Krstic, M. (1998). Stability margins in inverse optimal input-to-state stabilization. In *Proceedings of the 1998 American control conference. ACC (IEEE Cat. No. 98CH36207): 3* (pp. 1648–1652). IEEE.

Krstic, M., Kanellakopoulos, I., & Kokotovic, P.V. (1995). Nonlinear and adaptive control design.

Krstic, M., & Li, Z. H. (1998). Inverse optimal design of input-to-state stabilizing nonlinear controllers. *IEEE Transactions on Automatic Control, 43*(3), 336–350.

Krstic, M., & Tsiotras, P. (1997). Inverse optimality results for the attitude motion of a rigid spacecraft. In *Proceedings of the 1997 American control conference (Cat. No. 97CH36041): 3* (pp. 1884–1888). IEEE.

Krstic, M., & Tsiotras, P. (1999). Inverse optimal stabilization of a rigid spacecraft. *IEEE Transactions on Automatic Control, 44*(5), 1042–1049.

Kuhnen, K., & Janocha, H. (1999). Adaptive inverse control of piezoelectric actuators with hysteresis operators. In *1999 European control conference (ECC)* (pp. 791–796). IEEE.

Kurz, M. (1969). On the inverse optimal problem. In *Mathematical systems theory and economics i/ii* (pp. 189–201). Berlin, Heidelberg: Springer.

Lasserre, J. B. (2001). Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization, 11*(3), 796–817.

Lendaris, G. G., & Paintz, C. (1997). Training strategies for critic and action neural networks in dual heuristic programming method. In *Proceedings of international conference on neural networks (ICNN'97): 2* (pp. 712–717). IEEE.

Letov, A. M. (1960). Analytical design of controllers. *I. Avtom. Telemekh, 21*, 661–665.

Levine, S., & Koltun, V. (2012). Continuous inverse optimal control with locally optimal examples. arXiv preprint arXiv:1206.4617.

Levine, S., Popovic, Z., & Koltun, V. (2010). Feature construction for inverse reinforcement learning. In *Advances in neural information processing systems* (pp. 1342–1350).

Levine, S., Popovic, Z., & Koltun, V. (2011). Nonlinear inverse reinforcement learning with gaussian processes. In *Advances in neural information processing systems* (pp. 19–27).

Lewis, F. L., & Liu, D. (Eds.). (2013). *Reinforcement learning and approximate dynamic programming for feedback control* (17) (Eds.). John Wiley & Sons.

L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. *IEEE Access, 5*, 7776–7797.

Li, W., Todorov, E., & Liu, D. (2011). Inverse optimality design for biological movement systems. In *IFAC proceedings volumes: 44* (pp. 9662–9667).

Li, Y., Yao, Y., & Hu, X. (2018). Continuous-Time Inverse Quadratic Optimal Control Problem. arXiv preprint arXiv:1811.00129.

Liberzon, D., Sontag, E. D., & Wang, Y. (1999). On integral-input-to-state stabilization. In *Proceedings of the 1999 American control conference (Cat. No. 99CH36251): 3* (pp. 1598–1602). IEEE.

Lin, Y., & Sontag, E. D. (1991). A universal formula for stabilization with bounded controls. *Systems & Control Letters, 16*(6), 393–397.

Luo, W., Chu, Y. C., & Ling, K. V. (2005). Inverse optimal adaptive control for attitude tracking of spacecraft. *IEEE Transactions on Automatic Control, 50*(11), 1639–1654.

Madhavan, S. K., & Singh, S. N. (1991). Inverse trajectory control and zero dynamics sensitivity of an elastic manipulator. In *1991 American control conference* (pp. 1879–1884). IEEE.

Maeda, G. J., Neumann, G., Ewerton, M., Lioutikov, R., Kroemer, O., & Peters, J. (2017). Probabilistic movement primitives for coordination of multiple human–robot collaborative tasks. *Autonomous Robots, 41*(3), 593–612.

Magni, L., & Sepulchre, R. (1997). Stability margins of nonlinear receding-horizon control via inverse optimality. *Systems & Control Letters, 32*(4), 241–245.

Markovikj, D. (2014). *Deep apprenticeship learning for playing games (Doctoral dissertation*. University of Oxford).

Maruyama, A., & Fujita, M. (1999). Inverse optimal H∞ disturbance attenuation of robotic manipulators. In *1999 European control conference (ECC)* (pp. 2413–2418). IEEE.

Maslovskaya, S. (2018). *Inverse optimal control: Theoretical study (Doctoral dissertation*. Paris Saclay).

McShane, E. J. (1939). On multipliers for Lagrange problems. *American Journal of Mathematics, 61*(4), 809–819.

Mehdi, D., Darouach, M., & Zasadzinski, M. (1994). Discrete-time LQ design from the viewpoint of the inverse optimal regulator. *Optimal Control Applications and Methods, 15*(3), 205–213.

Menner, M., Worsnop, P., & Zeilinger, M.N. (2018). Predictive modeling by infinite-horizon constrained inverse optimal control with application to a human manipulation task. arXiv preprint arXiv:1812.11600.

Menner, M., Worsnop, P., & Zeilinger, M. N. (2019). Constrained inverse optimal control with application to a human manipulation task. *IEEE Transactions on Control Systems Technology*.

Mes, M. R., & Rivera, A. P. (2017). Approximate dynamic programming by practical examples. In *Markov decision processes in practice* (pp. 63–101). Cham: Springer.

Metelli, A. M., Pirotta, M., & Restelli, M. (2017). Compatible reward inverse reinforcement learning. In *Advances in neural information processing systems* (pp. 2050–2059).

Michini, B., Walsh, T. J., Agha-Mohammadi, A. A., & How, J. P. (2015). Bayesian nonparametric reward learning from demonstration. *IEEE Transactions on Robotics, 31*(2), 369–386.

Molinari, B. (1973). The stable regulator problem and its inverse. *IEEE Transactions on Automatic Control, 18*(5), 454–459.

Mombaur, K., Truong, A., & Laumond, J. P. (2010). From human to humanoid locomotion—An inverse optimal control approach. *Autonomous robots, 28*(3), 369–383.

Montgomery, J. F. (1999). *Learning helicopter control through teaching by showing*. University of Southern California.

Moylan, P.J., & Anderson, B.D. (1973). Nonlinear regulator theory and an inverse optimal control problem.

Murray, J. J., Cox, C. J., Lendaris, G. G., & Saeks, R. (2002). Adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 32*(2), 140–153.

Neittaanmäki, P., Rudnicki, M., Rudnicki, M., & Savini, A. (1996). *Inverse problems and optimal design in electricity and magnetism (No. 35)*. Oxford University Press.

Neu, G., & Szepesvári, C. (2007). Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Proceedings of the twenty-third conference on uncertainty in artificial intelligence* (pp. 295–302). AUAI Press.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. *Icml, 1*, 2.

Nguyen, Q. P., Low, B. K. H., & Jaillet, P. (2015). Inverse reinforcement learning with locally consistent reward functions. In *Advances in neural information processing systems* (pp. 1747–1755).

Obermayer, R. W., & Muckler, F. A. (1965). *On the inverse optimal control problem in manual control systems*: Vol. 208. National Aeronautics and Space Administration.

Ornelas, F., Sanchez, E. N., & Loukianov, A. G. (2010). Discrete-time inverse optimal control for nonlinear systems trajectory tracking. In *49th IEEE conference on decision and control (CDC)* (pp. 4813–4818). IEEE.

Ortega, R., Rodriguez, A., & Espinosa, G. (1990). Adaptive stabilization of non-linearizable systems under a matching assumption. In *1990 American control conference* (pp. 67–72). IEEE.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., & Peters, J. (2018). An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics, 7*(1–2), 1–179.

Osa, T., Sugita, N., & Mitsuishi, M. (2014). Online trajectory planning in dynamic environments for surgical task automation. In *Robotics: Science and systems* (pp. 1–9).

Osipchuk, M., Bharadwaj, S., & Mease, K. D. (1997). Achieving good performance in global attitude stabilization. In *Proceedings of the 1997 American control conference (Cat. No. 97CH36041): 3* (pp. 1889–1893). IEEE.

Paraschos, A., Daniel, C., Peters, J. R., & Neumann, G. (2013). Probabilistic movement primitives. In *Advances in neural information processing systems* (pp. 2616–2624).

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

Park, J., & Chung, W. K. (2000). Analytic nonlinear H/sub/spl infin//inverse-optimal control for Euler-Lagrange system. *IEEE Transactions on Robotics and Automation, 16*(6), 847–854.

Park, J. G., & Lee, K. Y. (1975). *An inverse optimal control problem and its application to the choice of performance index for economic stabilization policy: 1* (pp. 64–76). IEEE Transactions on Systems, Man, and Cybernetics.

Park, T., & Levine, S. (2013). Inverse optimal control for humanoid locomotion. *Robotics science and systems workshop on inverse optimal control and robotic learning from demonstration*.

Pathak, D., Mahmoudieh, P., Luo, G., Agrawal, P., Chen, D., & Shentu, Y. (2018). Zero-shot visual imitation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 2050–2053).

Pauwels, E., Henrion, D., & Lasserre, J. B. (2014). Inverse optimal control with polynomial optimization. In *53rd IEEE conference on decision and control* (pp. 5581–5586). IEEE.

Pauwels, E., Henrion, D., & Lasserre, J. B. (2016). Linear conic optimization for inverse optimal control. *SIAM Journal on Control and Optimization, 54*(3), 1798–1825.

Pirotta, M., & Restelli, M. (2016). Inverse reinforcement learning through policy gradient minimization. *Thirtieth AAAI conference on artificial intelligence*.

Plett, G. L. (2003). Adaptive inverse control of linear and nonlinear systems using dynamic neural networks. *IEEE Transactions on Neural Networks, 14*(2), 360–376.

Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., & Mishchenko, E.F. (1961). Mathematical Theory of Optimal Processes{in Russian}.

Porter, B., & Woodhead, M. A. (1970). Synthesis of an aircraft roll-stabilisation system: an application of invERSE OPTIMAL CONTROL THEORY. *The Aeronautical Journal, 74*(713), 390–392.

Powell, W. B. (2008). Approximate dynamic programming: Lessons from the field. In *2008 winter simulation conference* (pp. 205–214). IEEE.

Prasanna, P., Jacob, J., & Nandakumar, M. P. (2019). Inverse optimal control of a class of affine nonlinear systems. *Transactions of the Institute of Measurement and Control, 41*(9), 2637–2650.

Priess, M. C., Conway, R., Choi, J., Popovich, J. M., & Radcliffe, C. (2014). Solutions to the inverse lqr problem with application to biological systems analysis. *IEEE Transactions on Control Systems Technology, 23*(2), 770–777.

Puydupin-Jamin, A. S., Johnson, M., & Bretl, T. (2012). A convex approach to inverse optimal control and its application to modeling human locomotion. In *2012 IEEE international conference on robotics and automation* (pp. 531–536). IEEE.

Radoslav, S. (1988). On inverse problem of nonlinear system dynamics. In *Analysis and optimization of systems* (pp. 227–238). Berlin, Heidelberg: Springer.

Ramachandran, D., & Amir, E. (2007). In *Bayesian inverse reinforcement learning: 7* (pp. 2586–2591). IJCAI.

Ratliff, N., Bradley, D., Bagnell, J. A., & Chestnutt, J. (2006a). Boosting structured prediction for imitation learning. In *Proceedings of the 19th international conference on neural information processing systems* (pp. 1153–1160). MIT Press.

Ratliff, N. D., Bagnell, J. A., & Zinkevich, M. A. (2006b). *Maximum margin planning. In Proceedings of the 23rd international conference on machine learning* (pp. 729–736). ACM. 2006a.

Ratliff, N. D., Silver, D., & Bagnell, J. A. (2009). Learning to search: Functional gradient techniques for imitation learning. *Autonomous Robots, 27*(1), 25–53.

Ravanbakhsh, H., & Sankaranarayanan, S. (2019). Learning control Lyapunov functions from counterexamples and demonstrations. *Autonomous Robots, 43*(2), 275–307.

Ravi, S., & Larochelle, H. (2016). Optimization as a model for few-shot learning.

Rekasius, Z., & Hsia, T. (1964). On an inverse problem in optimal control. *IEEE Transactions on Automatic Control, 9*(4), 370–375.

Rohrweck, H., Schwarzgruber, T., & del Re, L. (2015). Approximate optimal control by inverse CLF approach. *IFAC-PapersOnLine, 48*(11), 286–291.

Ross, S., & Bagnell, D. (2010). Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 661–668).

Ross, S., Gordon, G., & Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 627–635).

Rouot, J., & Lasserre, J. B. (2017). On inverse optimal control via polynomial optimization. In *2017 IEEE 56th annual conference on decision and control (CDC)* (pp. 721–726). IEEE.

Russell, S. J. (1998). Learning agents for uncertain environments. *COLT, 98*, 101–103.

Saeks, R. E., Cox, C. J., Mathia, K., & Maren, A. J. (1997). Asymptotic dynamic programming: Preliminary concepts and results. In *Proceedings of international conference on neural networks (ICNN'97): 4* (pp. 2273–2278). IEEE.

Sanchez, E. N., & Ornelas-Tellez, F. (2017). *Discrete-time inverse optimal control for nonlinear systems*. CRC Press.

Sanchez, E. N., Perez, J. P., Martinez, M., & Chen, G. (2002). Chaos stabilization: An inverse optimal control approach. *Latin American Applied Research, 32*(1), 111–114.

Schaal, S. (1997). Learning from demonstration. In *Advances in neural information processing systems* (pp. 1040–1046).

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

Schaal, S., Ijspeert, A., & Billard, A. (2003). Computational approaches to motor learning by imitation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1431), 537–547.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning* (pp. 1889–1897).

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Schweitzer, P. J., & Seidmann, A. (1985). Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications, 110*(2), 568–582.

Sepulchre, R., Jankovic, M., & Kokotovic, P. (1997). Constructive Nonlinear Control.

Sepulchre, R., Jankovic, M., & Kokotovic, P. V. (2012). *Constructive nonlinear control*. Springer Science & Business Media.

Shahmansoorian, A. (2009). Inverse optimal control and construction of control Lyapunov functions. *Journal of Mathematical Sciences, 161*(2), 297–307.

Shiarlis, K., Messias, J., & Whiteson, S. (2016). Inverse reinforcement learning from failure. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems* (pp. 1060–1068). International Foundation for Autonomous Agents and Multiagent Systems.

Si, J., Barto, A. G., Powell, W. B., & Wunsch, D. (Eds.) (2004). *Handbook of learning and approximate dynamic programming* (2) (Eds.). John Wiley & Sons.

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

Sontag, E. D. (1989). A universal construction of Artstein's theorem on nonlinear stabilization. *Systems & control letters, 13*(2), 117–123.

Sontag, E. D. (1983). A Lyapunov-like characterization of asymptotic controllability. *SIAM Journal on Control and Optimization, 21*(3), 462–471.

Spong, M. W., & Ortega, R. (1990). On adaptive inverse dynamics control of rigid robots. *IEEE Transactions on Automatic Control, 35*(1), 92–95.

Sugimoto, K. (1998). Partial pole placement by LQ regulators: An inverse problem approach. *IEEE Transactions on Automatic Control, 43*(5), 706–708.

Sun, M., & Ma, X. (2019). Adversarial Imitation Learning from Incomplete Demonstrations. arXiv preprint arXiv:1905.12310.

Sussmann, H. J., & Willems, J. C. (1997). 300 years of optimal control: From the brachystochrone to the maximum principle. *IEEE Control Systems Magazine, 17*(3), 32–44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Syed, U., Bowling, M., & Schapire, R. E. (2008). Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on machine learning* (pp. 1032–1039). ACM.

Syed, U., & Schapire, R. E. (2008). A game-theoretic approach to apprenticeship learning. In *Advances in neural information processing systems* (pp. 1449–1456).

Takano, W., & Nakamura, Y. (2015). Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. *The International Journal of Robotics Research, 34*(10), 1314–1328.

Taskar, B., Chatalbashev, V., Koller, D., & Guestrin, C. (2005). Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on machine learning* (pp. 896–903). ACM.

Thau, F. (1967). On the inverse optimum control problem for a class of nonlinear autonomous systems. *IEEE Transactions on Automatic Control, 12*(6), 674–681.

Torabi, F., Warnell, G., & Stone, P. (2018). Behavioral cloning from observation. arXiv preprint arXiv:1805.01954.

Tucker, A., Gleave, A., & Russell, S. (2018). Inverse reinforcement learning for video games. arXiv preprint arXiv:1810.10593.

Uchibe, E. (2018). Model-free deep inverse reinforcement learning by logistic regression. *Neural Processing Letters, 47*(3), 891–905.

Ude, A., Atkeson, C. G., & Riley, M. (2004). Programming full-body movements for humanoid robots by observation. *Robotics and autonomous systems, 47*(2–3), 93–108.

Urbancic, T., & Bratko, I. (1993). Learning to control dynamic systems. In D. Michie (Ed.), *Machine learning and statistical classification* (ed.). Ellis-Horwood.

Van Den Berg, J., Miller, S., Duckworth, D., Hu, H., Wan, A., & Fu, X. Y. (2010). Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations. In *2010 IEEE international conference on robotics and automation* (pp. 2074–2081). IEEE.

Vega, C., & Alzate, R. (2014). Inverse optimal control on electric power conversion. In *2014 IEEE international autumn meeting on power, electronics and computing (ROPEC)* (pp. 1–5). IEEE.

Vito, E. D., Rosasco, L., Caponnetto, A., Giovannini, U. D., & Odone, F. (2005). Learning from examples as an inverse problem. *Journal of Machine Learning Research, 6*(May), 883–904.

Wang, F. Y., Zhang, H., & Liu, D. (2009). Adaptive dynamic programming: An introduction. *IEEE computational intelligence magazine, 4*(2), 39–47.

Wei, Y. J., & Shieh, L. S. (1979). Synthesis of optimal block controllers for multivariable control systems and its inverse optimal-control problem. In *Proceedings of the institution of electrical engineers: 126* (pp. 449–456). IET.

Werbos, P. (1977). *Advanced forecasting methods for global crisis warning and models of intelligence* (pp. 25–38). General System Yearbook.

Werbos, P. J. (1992). Approximate dynamic programming for real-time control and neural modeling. In D. A. White, & D. A. Sofge (Eds.), *Handbook of intelligent control: Neural, fuzzy and adaptive approaches* (Eds.). New York: Van Nostrand ch. 13.

Widrow, B. (1987). Adaptive inverse control. In *Adaptive systems in control and signal processing 1986* (pp. 1–5). Pergamon.

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001

Widrow, B., & Plett, G. L. (1996). Adaptive inverse control based on linear and nonlinear adaptive filtering. In *Proceedings of international workshop on neural networks for identification, control, robotics and signal/image processing* (pp. 30–38). IEEE.

Widrow, B., & Walach, E. (2008). *Adaptive inverse control, reissue edition: A signal processing approach*. John Wiley & Sons.

Willems, J., & Van De Voorde, H. (1977). Inverse optimal control problem for linear discrete-time systems. *Electronics Letters, 13*(17), 493.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks, 11*(7–8), 1317–1329.

Wulfmeier, M., Ondruska, P., & Posner, I. (2015). Maximum entropy deep inverse reinforcement learning. arXiv preprint arXiv:1507.04888.

Xi, Y., & Li, D. (2019). *Predictive control: Fundamentals and developments*. John Wiley & Sons.

Yaman, F., Yakhno, V. G., & Potthast, R. (2013). A survey on inverse problems for applied sciences. *Mathematical problems in engineering* 2013.

Yeh, W. W. G. (1986). Review of parameter identification procedures in groundwater hydrology: The inverse problem. *Water Resources Research, 22*(2), 95–108.

Yu, T., Finn, C., Xie, A., Dasari, S., Zhang, T., Abbeel, P. et al. (2018). One-shot imitation from observing humans via domain-adaptive meta-learning. arXiv preprint arXiv:1802.01557.

Zhang, H., Wang, Z., & Liu, D. (2004). Chaotifying fuzzy hyperbolic model using adaptive inverse optimal control approach. *International Journal of Bifurcation and Chaos, 14*(10), 3505–3517.

Zheng, J., Liu, S., & Ni, L. M. (2014). Robust bayesian inverse reinforcement learning with sparse behavior noise. *Twenty-eighth AAAI conference on artificial intelligence*.

Zhifei, S., & Joo, E. M. (2012). A review of inverse reinforcement learning theory and recent advances. In *2012 IEEE congress on evolutionary computation* (pp. 1–8). IEEE.

Zhu, Z., & Hu, H. (2018). Robot learning from demonstration in robotic assembly: A survey. *Robotics, 7*(2), 17.

Ziebart, B.D. (2010). Modeling purposeful adaptive behavior with the principle of maximum causal entropy (Doctoral dissertation, figshare).

Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. *Aaai, 8*, 1433–1438.

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

Nematollah Ab Azar was born in 1978 in Dashtestan, Bushehr province, Iran. He received B.S and M.S degrees in Electronics and Mechatronics, respectively. Currently, he is a Ph.D. candidate in Electrical Engineering/Control theory at Imam Khomeini International University (IKIU), Qazvin, Iran. His-research interest includes Optimal Control, Machine Learning, Optimization, Robotics, Mechatronics, and Image Processing.

Aref Shahmansoorian received the B.S and M.S degrees on electrical engineering from Tehran University in 1993 and 1996 respectively and received Ph.D. in Control theory from K. N. Toosi University of Technology (KNTU), Tehran, in 2005. Currently, he is with the group of electrical engineering, Imam Khomeini International University (IKIU), Qazvin, as an assistant professor. His research interest includes Nonlinear Control Systems, Optimal Control, Robust Control, and Multivariable control.

Mohsen Davoudi received his Ph.D. in Electrical Engineering from Polytechnic University of Milan (Politecnico di Milano), Milan, Italy, in 2011. Currently he is an assistant professor at Imam Khomeini International University (IKIU), Qazvin, Iran. His research interest includes Fuzzy Logic, Neural Networks, Computational Intelligence, and Expert systems.

Please cite this article as: N. Ab Azar, A. Shahmansoorian and M. Davoudi, From inverse optimal control to inverse reinforcement learning: A historical review, Annual Reviews in Control, https://doi.org/10.1016/j.arcontrol.2020.06.001