



Information-theoretic policy learning from partial observations with fully informed decision makers

Tom Lefebvre

Dynamic Design Lab (D2LAB) from the Departement of Electromechanical, Systems and Metal Engineering, Faculty of Engineering and Architecture, Ghent University, Technologiepark 131, Zwijnaarde 9052, Belgium

ARTICLE INFO

Article history:

Received 7 June 2022

Revised 19 October 2022

Accepted 23 October 2022

Available online 27 October 2022

Edited by Jiwen Lu

Keywords:

Information-theory

Hidden Markov models

Bayesian methods

Imitation learning

Markov decision processes

ABSTRACT

In this work we formulate and treat an extension of the Imitation from Observations problem. Imitation from Observations is a generalisation of the well-known Imitation Learning problem where state-only demonstrations are considered. In our treatment we extend the scope of Imitation from Observations to feature-only demonstrations which could arguably be described as partial observations. Therewith we mean that the full state of the decision makers is unknown and imitation must take place on the basis of a limited set of features. We set out for methods that extract an executable policy directly from those features which, in the literature, would be referred to as Behavioural Cloning methods. Our treatment combines elements from probability and information theory and draws connections with entropy regularized Markov Decision Processes.

© 2022 The Author(s). Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

1. Introduction

Imitation Learning (IL) refers to the process by which a student tries to learn how to execute a task by observing an experienced teacher demonstrate the task. Usually the student is allowed to collect a number of those demonstrations before attempting the task [1–3]. IL is known to be useful to treat hard control problems such as driving [4] and grasping [5–7].

Conventionally it is assumed that the student has gained access to demonstrations that include both the teacher's *states* $\{x_t\}_t$ (e.g. position, velocities) as well as its *actions* $\{u_t\}_t$ (e.g. forces). By definition this restrictive problem statement rules out a number of potentially useful teachers solely because said actions cannot be accessed. Though, even more restrictive is the assumption that the demonstrations must relate to the student performing the task rather than the teacher since this suggests that the states and actions must be those of the student and not those of the teacher (e.g. a human demonstrating a task by grabbing the end-effector whilst the robot collects measurements from its joints). The latter is known as the *embodiment mismatch* problem [8,9]. Alternatively one might imagine scenario's where the students has mere ac-

cess to a sequence of features, $\{z_t\}_t$, representative of the teacher's demonstration, and cannot access it's own states and actions.

To accommodate these flaws, a smaller number of studies has began to pose and treat IL problems with *state-only* demonstrations. Here clearly only the teacher's *states* can be accessed [10–13]. It is argued that the so called *Imitation from Observation* (IfO) paradigm offers a more natural way to consider learning from a teacher, and exhibits more similarity with the way many biological agents appear to approach imitation [9]. However, as is implied by the term *state-only* demonstrations, many recent studies that treat the IfO problem still assume full state observability, i.e. $z_t = x_t$. This setting is arguably evenly restrictive as full IL, given how rare a teacher and a student are that have the exact same state. Think of a human demonstrating a trajectory to a robot without grasping the robot.

In this work we consider the setting where only a limited set of representative features or partial observations are available. We refer to this problem as the Imitation learning from Partial Observations (IfPO) problem. In the context of IfPO, the goal is for the teacher and student to have the same effect on the environment, rather than the student doing the exact same thing as its teacher. In the setting of IfPO, the goal is for the student to show-case behaviour that appears similar to the behaviour displayed by the teacher to an *objective observer* that is merely interested in those features that were collected in the first place.

E-mail address: tom.lefebvre@ugent.be

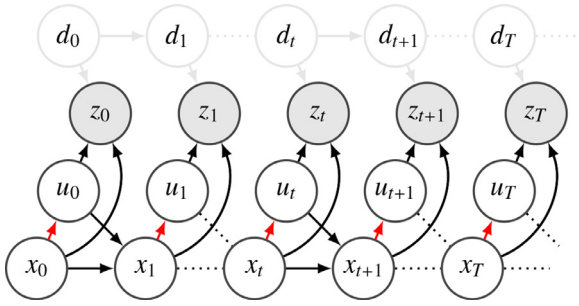


Fig. 1. Graphical representation of the imitation learning problem with partial observations and a fully informed decision maker.

Our methodological contributions are the following

1. First we formalize the IfPO problem and put forth a probabilistic model for how the student's dynamics may spawn the set of features displayed by the teacher.
2. We formulate a straightforward treatment of the IfPO problem by recasting it as a Bayesian inference problem which in turn leads us to derive a first learning algorithm.
3. Second, we recast the solution of the Bayesian inference as an information-theoretic projection. This reformulation allows us to generalise the treatment to multiple measurement sequences and to propose a sibling strategy by means of the reciprocal information-theoretic projection.

These aspects describe a first step towards a generic treatment of the IfPO problem and an extension to more general models. Then because the resulting policy learning algorithms demonstrate great similarity with the solution of Markov Decision Processes (MPDs), we pursue that intuition to some extent. Finally we specialize the concepts to linear-Gaussian systems. A main restriction of the proposed methods is the availability of the student's emission model (see later). Future work may focus on learning the emission model too but will likely require to embed the proposed approach in a more general framework.

2. Problem formulation

As a proxy for the student's imitation learning process, we propose the probabilistic graph model in Fig. 1

$$x_t \sim p(x_{t+1}|x_t, u_t) = \tau_t$$

$$u_t \sim p(u_t|x_t) = \rho_t$$

$$z_t \sim p(z_t|x_t, u_t) = \epsilon_t$$

Let us here briefly verbalize our notation. Any of the spaces $\mathcal{X}$, $\mathcal{U}$ or $\mathcal{Z}$ can be discrete or continuous. For brevity we introduce the variable tuple $\xi = (x, u)$ and the formatting $\underline{X}_t = \{x_0, \dots, x_t\}$ and $\bar{X}_t = \{x_t, \dots, x_T\}$ for leading or trailing subsequences. Throughout we use $\mathcal{P}$ to refer to sets of probability density functions. The set's arguments are implied by context.

1. $x_t \in \mathcal{X}$ represents the student's Markovian state, $\mathcal{X}$ denotes the state space. The probabilistic transition dynamics are governed by the (time-varying) functions $\{\tau_t\}_t$.
2. $u_t \in \mathcal{U}$ denotes the student's action, $\mathcal{U}$ denotes the action space. We assume that prior to the learning process the student has access to some auxiliary policy ρ_t .
3. $z_t \in \mathcal{Z}$ represent features of the student's shared with the teacher. $\mathcal{Z}$ denotes the feature space, the (time-varying) observation or emission model is given by $\{\epsilon_t\}_t$.
4. d_t represents the latent state of the teacher. The representation of the teacher's state is strictly demonstrative and

included as a justification for the origin of the feature sequence. However it is irrelevant what the teacher's exact state is and therefore also irrelevant what the teacher's transition and emission model are. Think of a human demonstrating a reference trajectory with its arm. We are unaware of the teacher's state nor its reasons to showcase that trajectory, yet we can measure the hand's trajectory.

We deliberately do not model the student's actions, $\{u_t\}_t$, to be the result of an MPD (see sec. 4.2). In general there is no reason to assume that the teacher acted rationally, i.e. acted so to optimize some utility function, so there is no reason for the student attempting to minimize some utility function. Furthermore this would imply that the student acts deterministically which we can also not know for sure. The main conceptual assumption is that the student and the teacher share the feature sequence $\underline{Z}_T$. So even though the feature sequence spawned from the dynamics of the teacher, we assume that with equal probability they could have spawned from the student. In this preliminary investigation, we also assume that the *proprioceptive* subsystem of the student is perfect and we have access to a perfect state estimate. Thus, the student's policy is governed by a probability density function conditioned on its state x_t . We refer to such a decision maker as being *fully informed*.

As such we consider the following problem

Problem 1. Learn policy functions $\{\pi_t\}_t$ provided that the decision maker is fully informed. Assume that the transition $\{\tau_t\}_t$ and emission $\{\epsilon_t\}_t$ models are known as well as multiple feature sequences $\{z_t^n\}_{t,n}$, representative for the dynamic behaviour that the decision maker's policy ought to induce.

By definition the strategies derived in this work are model-based. Our reasoning is that although the student might not know how to perform a task, it is reasonable to assume it has knowledge about its own transition model. For now we also assume that student has access to its emission model, i.e. a probabilistic mapping from state and actions to the relevant features. For example, the robot is aware of its own forward kinematics.

3. Algorithms for policy learning

We propose two, as far as we are aware of, original strategies to solve problem 1. In what follows we will refer to any of the proposed policy extraction strategies as an A Posteriori Control Distribution (APCD). Provided that we do not model the decision process as an MDP and that we assume to have access to a model, the strategies would classify as model-based BC algorithms. Though as mentioned earlier, the computational procedures are closely related to those characteristic for MDPs. The exact relation between the proposed BC algorithms and IRL methods will be investigated to some extent in Section 4.2. An important feature of the proposed methods is the ability to assimilate the information contained by multiple demonstrations.

The first strategy to derive an APCD follows a standard Bayesian argument. As a result, the concept only applies to a single measurement sequence. We investigate the solution's computational properties and whilst doing so explore whether the same result may be expressed as an optimization problem that generalizes over multiple feature sequences. The corresponding optimization problem turns out to be related to the information-theoretic Moment or M-projection. We refer to this APCD as the *vanilla* APCD (V-APCD). The terminology anticipates another APCD which we shall refer to as the *natural* APCD (N-APCD). The class of N-APCDs is obtained by considering and solving the reciprocal information-theoretic Information- or I-projection and is developed further in Section 3.3. Our use of terminology is inspired by *vanilla* and *natural*.

ural gradients in policy search where the natural gradient follows a relative entropy constraint corresponding the I-projection [14].

3.1. A primer on information-theoretic projections

Probabilistic inference refers to the process of reasoning with incomplete information according to rational principles. Inference principles therefore determine how to update a prior belief into a posterior belief when new information becomes available. Bayesian inference can be used to process information that is represented by the outcome of experiments, i.e. empirical evidence. The Information- and Moment-projection, respectively abbreviated to the I- and M-projection, are information-theoretic concepts that can be used to process information represented by constraints that affect the belief space. We could arguably refer to such evidence as *structural*.

The concepts are based on the relative entropy $\mathbb{D}[\pi \parallel \rho] = \mathbb{E}_\pi[\log \frac{\pi}{\rho}]$ between distributions $\pi \in \mathcal{P}$ and $\rho \in \mathcal{P}$. The relative entropy is a measure of the inefficiency of assuming that the distribution is ρ when the true distribution is π [15]. According to the principle of maximum entropy advocated by Jaynes amongst others, the relative entropy should be minimized if we want to encode some form of new information (usually an expectation, i.e. $\mathbb{E}_\pi[f] = \mu$) into the prior ρ .

- **I-projection** The I-projection and maximum entropy principle are equivalent with $\pi^\star$ the I-projection of ρ onto $\mathcal{P}^\star$.

$$\pi^\star = \arg \min_{\pi \in \mathcal{P}^\star} \mathbb{D}[\pi \parallel \rho]$$

- **M-projection** The M-projection is the reciprocal of the I-projection with $\pi^\star$ the M-projection of ρ onto $\mathcal{P}^\star$.

$$\pi^\star = \arg \min_{\pi \in \mathcal{P}^\star} \mathbb{D}[\rho \parallel \pi]$$

The set $\mathcal{P}^\star \subset \mathcal{P}$ represents the constrained belief space. The relative entropy is a divergence and not a distance and thus asymmetric in its arguments. Therefore the I-projection and the M-projection do not yield the same projection [16,17]. They are either *zero forcing* or *avoiding* for π . As a result the I-projection will underestimate the support of ρ and vice versa.

3.2. Vanilla APCD

First we describe the idea of the V-APCD for a single measurement sequence, $\underline{Z}_T$. Then we generalise the result to multiple measurement sequences, $\{\underline{Z}_T^n\}_n$, by reformulating the solution as an information-theoretic projection.

3.2.1. Bayesian argument

Reconsidering the interpretation of problem 1 we propose to synthesise the V-APCDs, $\{\pi_t^\star\}_t$, simply by conditioning the posterior marginal, $p(\xi_t | \underline{Z}_T)$, on the state, i.e. $p(u_t | x_t, \underline{Z}_T)$. This approach is closely related to BC algorithms, given that first we infer the most likely state-action distribution for given measurements and then infer the most likely policy. The ML community has developed various methods to practice inference on probabilistic graph models and hence the probabilities $p(\xi_t, \underline{Z}_T)$ and $p(u_t | x_t, \underline{Z}_T)$ could be calculated using one of these methods [18]. However such a general purpose method would not fully exploit the computational structure of the model nor would it yield further insights in the problem at hand.

$$\pi_t^\star(x_t) = p(u_t | x_t, \underline{Z}_T)$$

As a result of the Markov assumptions we have that the posterior marginal must be equivalent to $p(u_t | x_t, \underline{Z}_t)$. That is we can

drop the leading subsequence $\underline{Z}_{t-1}$ from $\underline{Z}_T$. This observation resonates with the common sense that once we arrive at some state, x_t , we can only hope to reproduce measurements $\underline{Z}_t$ but can no longer hope to affect the preceding measurements $\underline{Z}_{t-1}$. Thus we use Bayes' rule to decompose the V-APCD as follows.

$$p(u_t | x_t, \underline{Z}_t) = \rho_t(u_t | x_t) \frac{p(\underline{Z}_t | \xi_t)}{p(\underline{Z}_t | x_t)}$$

The decomposition reduces the problem to finding efficient expressions for the probabilities $p(\underline{Z}_t | x_t)$ and $p(\underline{Z}_t | \xi_t)$. The latter can be recognized as a generalisation of the backward filtering distribution. The former can be derived from there.

$$p(\underline{Z}_t | x_t) = \int \rho_t(u_t | x_t) p(\underline{Z}_t | \xi_t) du_t = \mathbb{E}_{\rho_t(u_t | x_t)} [p(\underline{Z}_t | \xi_t)]$$

One now easily verifies that the distribution $p(\underline{Z}_t | \xi_t)$ is governed by a backward recursive expression.

$$\begin{aligned} p(\underline{Z}_t | \xi_t) &= p(z_t | \xi_t) \int \tau_t(x_{t+1} | \xi_t) p(\underline{Z}_{t+1} | x_{t+1}) dx_{t+1} \\ &= p(z_t | \xi_t) \mathbb{E}_{\tau_t(x_{t+1} | \xi_t)} [p(\underline{Z}_{t+1} | x_{t+1})] \end{aligned}$$

Both the problem statements themselves, as well as the backward recursive calculation procedure, hint at a connection with the theory of dynamic programming. To illuminate the connection we may define the following negative log-probabilities, where we used notation $l(\cdot) \equiv -\log p(\cdot)$.

$$\begin{aligned} Q_t^\star(\xi_t) &= l(\underline{Z}_t | \xi_t) = -\log \epsilon_t(z_t | \xi_t) - \log \mathbb{E}_{\tau_t(x' | \xi_t)} [\exp(-V_{t+1}^\star(x'))] \\ V_t^\star(x_t) &= l(\underline{Z}_t | x_t) = -\log \mathbb{E}_{\rho_t(u_t | x_t)} [\exp(-Q_t^\star(\xi_t))] \end{aligned} \quad (1)$$

whereas the V-APCD is given by

$$\pi_t^\star(u_t | x_t) = \rho_t(u_t | x_t) \exp(V_t^\star(x_t) - Q_t^\star(\xi_t)) \quad (2)$$

3.2.2. Information-theoretic argument

Here we raise the question whether the V-APCD renders some objective function optimal? In particular we show that the V-APCD is governed by the M-projection of $p(\underline{\Xi}_T | \underline{Z}_T)$ onto the probability space spanned by $p(\underline{\Xi}_T; \underline{\pi}_T)$. Here $p(\underline{\Xi}_T; \underline{\pi}_T)$ is defined as the joint distribution obtained by administering some probabilistic policy sequence $\underline{\pi}_T$ instead of the assumed probabilistic control model $\rho_T(\underline{U}_T | \underline{X}_T)$ ¹. Another way of looking at this is that we want to identify a probabilistic control model that induces a trajectory distribution that discriminates less from the posterior trajectory distribution. Casting the V-APCD as an optimization problem also allows to generalise the concept to multiple measurement sequences $\{\underline{Z}_T^n\}$. Assuming that the sequences are i.i.d. we propose to sum their contributions.

As such we consider following variational optimization problem that can be manipulated into $T-1$ separate subproblems.

$$\begin{aligned} \arg \min_{\underline{\pi}_T \in \mathcal{P}} \sum_n \mathbb{D}[p(\underline{\Xi}_T | \underline{Z}_T^n) \parallel p(\underline{\Xi}_T; \underline{\pi}_T)] \\ = \arg \max_{\underline{\pi}_T \in \mathcal{P}} \sum_n \int p(\underline{\Xi}_T | \underline{Z}_T^n) \log \frac{p(\underline{\Xi}_T; \underline{\pi}_T)}{p(\underline{\Xi}_T | \underline{Z}_T^n)} d\underline{\Xi}_T \\ = \arg \min_{\underline{\pi}_T \in \mathcal{P}} \sum_t \sum_n \int p(\underline{\Xi}_T | \underline{Z}_T^n) \log \pi_t(u_t | x_t) d\underline{\Xi}_T \end{aligned} \quad (3)$$

The solution to this problem is governed by the V-APCD that we derived earlier for $N=1$ and by a mixture of individual V-APCDs for $N>1$. A derivation is provided in Appendix A (see supplementary material of the web version of this article). The mixture weights are determined by the individual smoothing distri-

¹ Note that therefore $p(\underline{\Xi}_T | \underline{Z}_T)$ is parametrized by the prior probabilistic control model ρ_T . To keep notation light this subtlety was not included.

butions and can be calculated using traditional smoothing algorithms substituting the closed loop transition and emission models, $\tau'_t = \mathbb{E}_{\rho_t}[T_t]$ and $\epsilon'_t = \mathbb{E}_{\rho_t}[E_t]$. Thus the ensemble V-APCD determines how much each individual V-APCD contributes based on the probability that the would occupy the state, x_t , for each of the individual N feature sequences $\bar{Z}_t^n$.

$$\pi_t^*(u_t|x_t) = \frac{\sum_n p(\xi_t|\bar{Z}_t^n)}{\sum_n p(x_t|\bar{Z}_t^n)} = \sum_n \frac{p(x_t|\bar{Z}_t^n)}{\sum_n p(x_t|\bar{Z}_t^n)} p(u_t|x_t, \bar{Z}_t^n) \quad (4)$$

3.3. Natural APCD

Reasoning from the information-theoretic motivation of the V-APCD above, we raise now the obvious question whether the reciprocal I-projection also generates an APCD? According to the information-theoretic interpretation of the I-projection it follows that therefore we minimize the inefficiency of assuming the prior probabilistic control model ρ_T whilst the true probabilistic control model is given by the posterior π_T . This generates a sibling APCD. Provided our earlier discussion on terminology, we refer to the APCD derived here as the *natural* APCD (N-APCD). We generalise the concept to multiple measurement sequences by superposing the individual contributions.

Thus we consider the variational optimization problem

$$\min_{\pi_T \in \mathcal{P}} \sum_n \mathbb{D}[p(\Xi_T; \pi_T) \parallel p(\Xi_T|\bar{Z}_T^n)] \quad (5)$$

This problem can be recast as follows (Appendix B)

$$\min_{\pi_T \in \mathcal{P}} \mathbb{E}_{p(\Xi_T; \pi_T)} \left[-\frac{1}{N} \sum_n \log p(\bar{Z}_T^n|\Xi_T) + \log \frac{\pi_T(U_T|X_T)}{\rho_T(U_T|X_T)} \right] \quad (6)$$

The problem above exhibits an optimal substructure which permits application of the principle of dynamic programming.

In particular we can decompose the optimization problem as demonstrated below

$$\min_{\pi_T \in \mathcal{P}} \int p(\Xi_t, x_{t+1}; \pi_T) \times \left(-\frac{1}{N} \sum_n \log p(\bar{Z}_t^n|\Xi_t) + \log \frac{\pi_t(U_t|X_t)}{\rho_t(U_t|X_t)} \right) d\Xi_t V_{t+1}^*(x_{t+1}) dx_{t+1}$$

Here we have defined the following value function

$$V_t^*(x_t) = \min_{\pi_t \in \mathcal{P}} \mathbb{E}_{p(\Xi_t|x_t, \pi_t)} \left[-\frac{1}{N} \sum_n \log p(\bar{Z}_t^n|\Xi_t) + \log \frac{\pi_t(\bar{U}_t|\bar{X}_t)}{\rho_t(\bar{U}_t|\bar{X}_t)} \right]$$

Inspired by the structural connection with MDPs that was already hinted at in [sec. 3.2.1](#), one easily verifies that also V_t^* satisfies a backward recursive calculation procedure

$$V_t^*(x_t) = \min_{\pi_t \in \mathcal{P}} \mathbb{E}_{\pi_t(u_t|x_t)} \left[\log \frac{\pi_t(x_t|u_t)}{\rho_t(u_t|x_t)} + Q_t^*(\xi_t) \right]$$

where

$$Q_t^*(\xi_t) = -\frac{1}{N} \sum_n \log \epsilon_t(z_t^n|\xi_t) + \mathbb{E}_{\tau_t(x_{t+1}|\xi_t)} [V_{t+1}^*(x_{t+1})] \quad (7)$$

Variational optimization (similar to Appendix A) of this final problem then yields an explicit expression for the value function V_t^* and thus the desired N-APCD, π_t^*

$$\begin{aligned} V_t^*(x_t) &= -\log \mathbb{E}_{\rho_t(u_t|x_t)} [\exp(-Q_t^*(\xi_t))] \\ \pi_t^*(x_t|u_t) &= \rho_t(u_t|x_t) \exp(V_t^*(x_t) - Q_t^*(\xi_t)) \end{aligned} \quad (8)$$

3.4. Some first observations

We present here a number of observations regarding the methodologies set-out so far.

1. On account of the underlying probabilistic graph model, either APCD are governed by a *Bayesian type* update rule. The probabilistic control model, ρ_t , acts as a prior probability which is updated to a posterior probabilistic control model, π_t , by the likelihood function, $\exp(-Q_t^*)$, $* \in \{\star, \bullet\}$.
2. For $N=1$, the solution of the *vanilla* and *natural* problems appear equivalent. At least, so do the backwards recursive expressions for the associated policy distributions and value functions. The difference lies in the definition of the Q -functions. Comparing the expressions derived for Q_t^* (1) and $Q_t^\bullet$ (7) reveals that the former additionally transforms the expectation in agreement with the probability-likelihood transformation (i.e. $l \equiv -\log p$) so that the apparent addition in likelihood space in fact amounts to a multiplication in probability space. In contrast, the *natural* approach is to carry out the computation in likelihood space. Further note that the difference is rendered irrelevant for deterministic dynamics in which case $Q_t^* \equiv Q_t^\bullet$ and therefore $\pi_t^* \equiv \pi_t^\bullet$ and so is $V_t^* \equiv V_t^\bullet$.
3. We can further analyse the difference in dynamics induced by both APCDs by considering the following decomposition of the posterior density, $p(\Xi_T|\bar{Z}_T)$. We obtain

$$p(\Xi_T|\bar{Z}_T) = p(x_0|\bar{Z}_T) p(u_0|x_0, \bar{Z}_T) \prod_t p(\xi_t|\xi_{t-1}, \bar{Z}_t)$$

where

$$p(\xi_t|\xi_{t-1}, \bar{Z}_t) = p(x_t|\xi_{t-1}, \bar{Z}_t) p(u_t|x_t, \bar{Z}_t)$$

This decomposition thus serves as an a posteriori justification of the Bayesian argument in [3.2.1](#). Unfortunately it also illustrates that the V-APCD will only induce the true posterior trajectory distribution $p(\Xi_T|\bar{Z}_T)$ if the system were governed by the *informed* transition probability $p(x_t|\xi_{t-1}, \bar{Z}_t)$. Clearly when the student practices the V-APCD, it cannot influence its own inherent dynamics so that we must substitute the natural transition probability rather than the informed transition probability. This means that the probability $p(\Xi_T; \pi_T^*)$ differs from the probability $p(\Xi_T|\bar{Z}_T)$. The reason is that the V-APCD aims to reconstruct the conditional distribution $p(u_t|x_t, \bar{Z}_t)$ and not the trajectory distribution itself. The V-APCD is naive in that sense. A similar analysis is not possible for the N-APCD, though it is anticipated that the N-APCD tries to accommodate for the difference between the natural and informed transition probabilities.

3.5. Linear-Gaussian systems

Here we specialize the APCDs to Linear-Gaussian (LG) systems

$$\begin{aligned} x_{t+1} &\sim T_t = \mathcal{N}(x_{t+1}; F_{\xi,t} \xi_t + f_t, Q_t) \\ u_t &\sim \rho_t = \mathcal{N}(u_t; K_t x_t + k_t, \Sigma_t) \\ z_t &\sim E_t = \mathcal{N}(z_t; G_{\xi,t} \xi_t + g_t, R_t) \end{aligned}$$

In this setting it is can be anticipated that the APCDs behave as affine Gaussian probabilities $\pi_t^* = \mathcal{N}(u_t|K_t^* x_t + k_t^*, \Sigma_t^*)$ and that both Q_t^* and V_t^* will be quadratic in their arguments. Computational details are given in Appendix C. Updates are given for $N=1$. Extension to (8) is trivial given (7). Extension to (4) requires calculating the smoothing distributions, $p(x_t|\bar{Z}_t^n)$ [18].

With $* \in \{\star, \bullet\}$ throughout, the policy parameters are given

$$\begin{aligned} k_t^* &= \Sigma_t^* (S_t^{-1} k_t - Q_{u,t}^*) \\ K_t^* &= \Sigma_t^* (S_t^{-1} K_t - Q_{ux,t}^*) \\ \Sigma_t^* &= (S_t^{-1} + Q_{uu,t}^*)^{-1} \end{aligned} \quad (9)$$

where $Q_{\xi,t}^*$ and $Q_{\xi,t}^*$ parametrise the quadratic model for Q_t^* . The parameters can be calculated recursively using the following expressions. Parameters $r_{\xi,t}$ and $r_{\xi,t}^*$ relate to the quadratic model of the negative logarithm of the emission model.

$$\begin{aligned} Q_{\xi,t}^* &= r_{\xi,t} + F_{\xi,t}^\top (V_{xx,t+1}^{*-1} + \mathbf{1}_*(*) \cdot Q_t)^{-1} (V_{xx,t+1}^{*-1} V_{x,t}^* + f_t) \\ Q_{\xi,t}^* &= r_{\xi,t}^* + F_{\xi,t}^\top (V_{xx,t+1}^{*-1} + \mathbf{1}_*(*) \cdot Q_t)^{-1} F_{\xi,t} \end{aligned} \quad (10)$$

Similarly $V_{\xi,t}^*$ and $V_{\xi,t}^*$ are parameters from the quadratic model for V_t^* and subject to the following recursions.

$$\begin{aligned} V_{x,t}^* &= Q_{x,t}^* + K_t^\top S_t^{-1} K_t - K_t^{*\top} \Sigma_t^{*-1} K_t^* \\ V_{xx,t}^* &= Q_{xx,t}^* + K_t^\top S_t^{-1} K_t - K_t^{*\top} \Sigma_t^{*-1} K_t^* \end{aligned} \quad (11)$$

Practical procedures are given in Algorithm 1 and 2.

Algorithm 1 LG-V-APCD.

Require: $\{\{F_{\xi,t}, f_t, Q_t\}\}_t, \{\{K_t, k_t, S_t\}\}_t, \{\{G_{\xi,t}, g_t, R_t\}\}_t, \{z_t^n\}_{t,n}$

Ensure: $\{\pi_t^*\}_t$

```

1: for  $n \in \{1, 2, \dots, N\}$  do
2:    $p(\xi_t | Z_t^n) = \mathcal{N}(\mu_{\xi,t}, \Sigma_{\xi,t}, \forall t \in 0, 1, \dots, T$  [18]
3:   for  $t \in \{0, 1, \dots, T-1\}$  do
4:      $\mathcal{N}(u_t; k_t^{n,*} + K_t^{n,*} x_t, \Sigma_t^{n,*}) = p(u_t | x_t, Z_t^n)$  [19]
5:     compute  $\pi_t^*$  according to (4)
6:   end for
7: end for
```

Algorithm 2 LG-N-APCD.

Require: $\{\{F_{\xi,t}, f_t, Q_t\}\}_t, \{\{K_t, k_t, S_t\}\}_t, \{\{G_{\xi,t}, g_t, R_t\}\}_t, \{z_t^n\}_{t,n}$

Ensure: $\{\pi_t^*\}_t$

```

1: for  $t \in \{T, T-1, \dots, 1\}$  do
2:   update  $\{V_{x,t}^*, V_{xx,t}^*\}$  according to (11)
3:   update  $\{Q_{\xi,t-1}^*, Q_{\xi,t-1}^*\}$  according to (10)
4:   update  $\{k_{t-1}^*, K_{t-1}^*, \Sigma_{t-1}^*\}$  according to (9)
5: end for
```

4. Related work

4.1. Imitation learning (from observations)

The goal of IL is to find a time-invariant policy function, $\pi : x \mapsto u$, (mapping states to actions), or, a sequence of time-varying policy functions, $\{\pi_t\}$, so that the closed-loop dynamics of the student produces behaviour similar to that of the teacher.

There are roughly two dominant approaches to face this problem, *behavioural cloning* (BC) and *inverse reinforcement learning* (IRL) [2,9,19]. IRL methods model the behaviour of the student and teacher as a Markov Decision Process (MDP) and try to infer the cost/reward function that is used by the teacher to make policy in the belief that it is the most concise and portable representation of the task [1,20]. IRL offers insight as to why the teacher makes certain decisions, though the often time consuming policy (re)construction is delayed to post-processing.

BC is powerful in the sense that it requires only demonstration data to directly learn an imitation policy and does not require any further interaction between the agent and the environment. A natural approach to BC directly targets the mapping from states to actions through supervised learning. Learning of the policy π reduces to solving the following problem. Recent advances in IL focus on the infinite horizon setting, $p(x, u; \mathcal{D})$ thus represents the stationary state-action data distribution.

$$J_{BC}[\pi; \mathcal{D}] = \mathbb{E}_{p(x,u;\mathcal{D})} [\log \pi(u|x)]$$

Adversarial IL (AIL) methods have shown great success in benchmarks for continuous control, especially in the low data regime [19,21]. AIL directly aim to recover the policy similar to BC, yet are closely related to the MDP formulation of IRL. As it turns out many existing IL methods can be unified as IL by f -divergence minimization, so called *distribution matching* [21,22]. The reverse minimization is mode-seeking and preferable.

$$J_{AIL}[\pi; \mathcal{D}] = \begin{cases} \mathbb{D}_f[p(x, u; \mathcal{D}) || p(x, u; \pi)], & \text{forward} \\ \mathbb{D}_f[p(x, u; \pi) || p(x, u; \mathcal{D})], & \text{reverse} \end{cases}$$

As stated in the introduction, IfO relaxes the requirement on action labels, and aims to imitate the expert's behaviour only from the state observations. The amount of studies that focus on the IfO problem remains however limited [9,11,22–26]. A natural approach is to mimic BC by augmenting the state-only demonstrations with action labels. Roughly summarized inverse dynamic approaches invert consecutive states, $\{x_t, x_{t+1}\}$, into an action, u_t , and then progress along the direction of standard BC approaches by supervised learning [25,27–29]. The reconstruction loss may range from simple least-squares regression to more complex losses such as e.g. inverse dynamics disagreement. More recently the concept of distribution matching has also been adopted. However, since the classical approach is no longer applicable, the distribution matching is recast in terms of the stationary state-transition distribution [22,24–26]. A unified view is given below. The hyperparameter $\alpha > 0$ balances between encouraging state-transition matching and preventing distribution shift from a set of imperfect demonstrations, $\mathcal{U}$.

$$J_{IfO}[\pi; \mathcal{D}] = (1 - \alpha) \mathbb{D}[p(x, x'; \pi) || p(x, x'; \mathcal{D})] + \alpha \mathbb{D}[p(x, u; \pi) || p(x, u; \mathcal{U})]$$

In light of the divergence minimization or distribution matching frameworks tailored to either IL and IfO reviewed above, it is interesting to revise the APCDs objectives from this work

$$J_{IfPO}[\pi; \mathcal{D}, \rho] = \begin{cases} \sum_n \mathbb{D}[p(\Xi_T | Z_T^n; \rho_T) || p(\Xi_T; \pi_T)], & \text{V-APCD} \\ \sum_n \mathbb{D}[p(\Xi_T; \pi_T) || p(\Xi_T | Z_T^n; \rho_T)], & \text{N-APCD} \end{cases}$$

4.2. Connections with entropy regularized MDPs

Although we have deliberately not modelled the decision process of the student as an MDP, there have been indirect suggestions that either APCD solve some sort of Optimal Control problem. In order to make that intuition explicit, let us recall the theory of MDPs. An MDP is defined as follows

$$\arg \min_{\bar{U}_0} \mathbb{E}_{p(\bar{X}_1 | x_0, \bar{U}_0)} [R_T(\bar{\Xi}_T)]$$

The objective function is given by the cumulative cost $R_T(\bar{\Xi}_T) = \sum_t r_t(\xi_t)$. The solution is governed by a deterministic feedback policy sequence, $\{\pi_t^*\}_t$. Similar to the probabilistic APCDs, the sequence is governed by a backward recursion (12). Substituting $\mathbb{E}_{\pi_t}[Q_t^*(\xi_t)] + \mathbb{D}[\pi_t || \rho_t]$ for $Q_t^*(\xi_t)$ and $l(z_t | \xi_t)$ for $r_t(\xi_t)$, we retrieve the optimization problem in (7).

$$\begin{aligned} \pi_t^*(x_t) &= \arg \min_{u_t \in \mathcal{U}} Q_t^*(\xi_t) \\ V_t^*(x_t) &= \min_{u_t \in \mathcal{U}} Q_t^*(\xi_t) \\ Q_t^*(\xi_t) &= r_t(\xi_t) + \mathbb{E}_{p(x' | \xi_t)} [V_{t+1}^*(x')] \end{aligned} \quad (12)$$

A risk-sensitive generalisation of the standard MDP has been developed where instead of minimizing the cumulative performance criteria $R_T(\Xi_T)$, the exponential of that objective, $\exp(-R_T(\Xi_T))$, is maximized [30]. This choice puts more emphasis on the contribution of the tails of the trajectory distribution than is done when using the expected cumulative cost function.

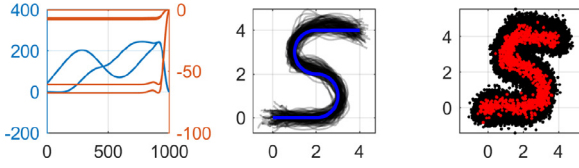


Fig. 2. Illustrative linear problem. We want to reconstruct the behavioural features of the force control feedback policy from position measurements. **Left:** feed-forward terms (blue) and feedback gains (red). **Middle:** 10^2 tracking experiments (black) about a reference path (blue). **Right:** 10^2 measurement sequences (black) with highlighted first measurement sequence (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The solution of the risk-sensitive MDP is also given by a deterministic policy sequence, $\{\pi_t^\nabla\}_t$, governed by the recursive calculation in (13). Upon execution of the same set of substitutions that we introduced in the setting of standard MDPs, one verifies that we retrieve the same problem statement as in (1).

$$\begin{aligned} p_t^\nabla(x_t) &= \arg \min_{u_t} Q_t^\nabla(\xi_t) \\ V_t^\nabla(x_t) &= \min_{u_t} Q_t^\nabla(\xi_t) \\ Q_t^\nabla(\xi_t) &= r_t(\xi_t) - \log \mathbb{E}_{p(x'|\xi_t)}[\exp(-V_{t+1}^\nabla(x'))] \end{aligned} \quad (13)$$

These observations suggest that there is a very strong connection between the MDP frameworks detailed here and the APCDs derived in sections 3.2 and 3.3, hence implying a strong connection between the BC approach that we originally set out for and IRL strategies that try to infer a cost model which is used consequently as input to solve an MDP. For a single measurement sequence we conclude that the N-APCD behaves as an entropy regularized MPD, and the V-APCD behaves as an entropy regularized risk-sensitive MDP, using the conditional measurement negative log-likelihood, $l(\underline{Z}_T|\underline{\Xi}_T)$, as cumulative cost. Intuitively it is sensible to substitute $l(\underline{Z}_T|\underline{\Xi}_T)$ as a proxy for $R_T(\underline{\Xi}_T)$ representing how likely it is to have traversed some trajectory $\underline{\Xi}_T$ given observation of the features $\underline{Z}_T$ similar to how eager we are to traverse that trajectory when we try to minimize $R_T(\underline{\Xi}_T)$. As a result of the entropy regularization we obtain expectation- rather than optimization operators and consequently we retrieve probabilistic policies rather than deterministic policies. In conclusion we note that the connection with the V-APCD breaks down for multiple sequences given that the solution is then given as a mixture of individual V-APCDs. For the N-APCDs the connection is maintained with the cumulative cost averaged out over the multiple demonstrations.

5. Experiments

In this section we document numerical experiments to validate the APCDs proposed in Sections 3.2 and 3.3. With our numerical experiments we want to clear out which of the distributions is preferable, $p(\underline{\Xi}_T; \pi_T^\nabla)$ or $p(\underline{\Xi}_T; \pi_T^\nabla)$. All experiments were implemented using Matlab. Each experiment was executed on a single 2.10GHz Intel Xeon Gold 6130 processor.

5.1. Problem definition

We consider a force controlled planar mass with Brownian input noise, see Fig. 2. The covariance of the input noise is spawned cascading the `rand` and `sprandsym` command generating correlated white noise. To realise anisotropy, the random seed is multiplied with $\text{diag}([10, 10^2])$. The system is discretised using a sample period $\Delta t = 2 \cdot 10^{-3}$ s. We define the teacher as a path tracking problem over the horizon $T = 2$ s. The teacher's policy is given by a Linear Quadratic Exponential Regulator (LQER) [30] minimizing the objective defined below. Here p_t , p_t^* and v_t define the position, desired position and velocity of the particle respectively.

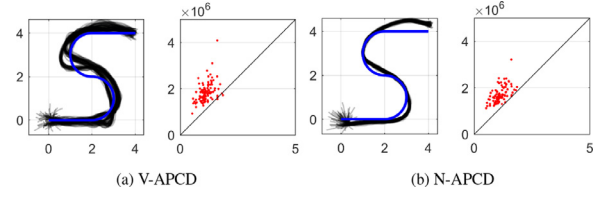


Fig. 3. Comparison of the V- and N-APCD for $\sigma^2 = 10^4$ and $N = 3$. The figures on the right compares the cost obtained with the learned student's policy (vertical) and the cost obtained with the teacher's policy (horizontal) when repeating the experiment with the same noise values. For experiments on the diagonal the same cost was obtained. For experiments above the diagonal, the teacher outperforms the student.

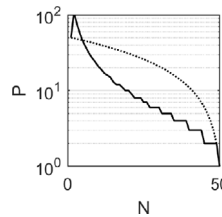
We set $W_p = 10^4 \cdot I$, $W_p = W_u = 1$ and $\lambda = 10^{-4}$. It is well known that the solution is given by a time dependent linear policy, i.e. $u_t = k_t^\square + K_t^\square x_t$. When simulating the system we use exact state observations though the system is tracked using position measurements, i.e. $z_t = p_t$, for post-processing. The covariance of the measurement noise was determined using the same procedure as the input noise, though here the random seed was multiplied with 10^{-1} in all dimensions. Finally we assume that the system is initialised with zero mean white noise with $\sigma_0^2 = 10^{-1}$.

$$\min \mathbb{E} \left[\exp \left(\frac{\lambda}{2} \sum_{t=0}^T \|p_t - p_t^*\|_{W_p}^2 + \|v_t\|_{W_v}^2 + \|u_t\|_{W_u}^2 \right) \right]$$

5.2. Results

We verify the capacity of the V-APCD and N-APCD to reconstruct the underlying policy $\{k_T^\square, K_T^\square\}$ using the APCD mean as a proxy for the LQER. Thus we interpret the APCD covariance as a measure for our epistemic uncertainty about the APCD mean. For the N-APCD this results into a linear feedback policy. For the V-APCD this results into a Gaussian mixture of N linear feedback policies. The policy prior, ρ_T , is characterised as a linear Gaussian policy with zero mean and uncorrelated covariance with magnitude σ^2 . Because we are interested in behavioural features of dynamics induced by the learned student's policy, rather than in a perfect reconstruction of the teacher's policy we do not quantify the reconstruction of the LQER itself but compare the performance of the APCDs with respect to the control objective defined above. In our experiments we vary two hyperparameters, in particular the magnitude σ^2 and the number of measurement sequences N picked randomly from half of $M = 10^2$ individual experiments. The APCDs are then validated on the same 10^2 experiments using the same in- and output noise. To counteract the influence of the specific N from half the M sequences on the reconstruction, we verify P unique but random combinations so that the probability of never having included a specific sequence is less than 1%² See the inline figure.

Fig. 3 visualizes the performance of both the V- and N-APCD respectively for $\sigma^2 = 10^4$ and $N = 3$. It is interesting to note that either APCD is capable of reconstructing the behavioural features of the LQER to visual satisfaction. One can verify that the V-APCD acts as a combination of 3 individual policies where the



² The number of unique combinations when picking N from M is $B = \binom{M}{N}$. The number of unique combinations that does not include a specific experiment is $A = \binom{M-1}{N}$. The probability of picking a combination that does not include a specific experiment on the $(p-1)$ th try equals $q(p) = \frac{A-p}{B-p}$. The probability of picking P unique combinations that do not include a specific sequence equals $f(P) = \prod_{p=1}^P q(p)$. As such we can compute P so that $f(P) \leq \bar{f}$ for given N .

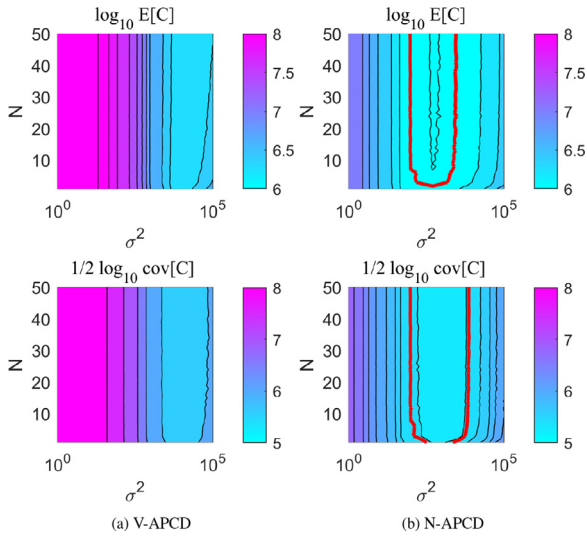


Fig. 4. Comparison of the V- and N-APCD performance as measured by the original control objective for varying σ^2 and N according to the principles described in Section 5.2. The red curves indicate the performance of the LQER for $M = 100$. Note that the horizontal scale is logarithmic. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

acting policy is determined by the measurement sequences that best explains the current state according to $p(x_t | Z_t^n)$. On the contrary the N-APCD averages out the contribution of each measurement sequence in likelihood space. As can be seen, for smaller N this results into a slight misalignment of the reconstructed desired behaviour and the true reference path. Fig. 4 documents results for varying σ^2 and N . Depending on N the experiment was repeated P times with a unique combination of sequences. Overall the N-APCDs can be seen to outperform the V-APCDs with the N-APCDs obtaining similar performance to the LQER.

6. Conclusion

In this work we formalized the problem of Imitation from Partial Observations and discussed two novel strategies to treat the Imitation from Observations problem from a single or multiple partial observations sequences. We have set-up a Behavioural Cloning approach based on the probabilistic graph model in Fig. 1. By reformulating the problem as an information-theoretic Moment projection, the strategy could be generalised to multiple observations sequences. Treatment of the reciprocal Information projection yielded a sibling solution to the imitation problem. Comparison with recent work on distribution matching tailored to Imitation Learning (from Observations) illustrates that our work can be classified as a distribution matching method for Imitation from Partial Observations.

Further investigation illustrated that both policy learning methods can also be interpreted as a form of Inverse Reinforcement Learning using a specific choice for the cost model and using entropy regularization to bias the inferred policy on some prior policy probability. Provided that the policies can also be seen as specific instances of entropy regularized Markov Decision Processes, it should be possible to add auxiliary features to the student's policy by extending the cost model with additional terms that reflect specific behaviour.

Specialization of our results to Linear-Gaussian models provided an explicit backward recursive calculation procedure to infer the desired policy distributions which was verified empirically. Preliminary sufficient conditions for which these procedures will yield feasible solutions were given leaning on classical estimation and

control theory. A more detailed analysis is still required to determine necessary conditions.

In future work it would also be interesting to extend the treatment to nonlinear probabilistic state-space models as well and to further explore the theoretical connection with the generic representation learning problem where neither the transition and/or the emission model are known.

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.

Acknowledgements

The authors wish to acknowledge financial support from the Research Foundation – Flanders (FWO), grant no. S007723N.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at [10.1016/j.patrec.2022.10.025](https://doi.org/10.1016/j.patrec.2022.10.025)

References

- [1] B.D. Argall, S. Chernova, M. Veloso, B. Browning, A survey of robot learning from demonstration, *Rob Auton Syst* 57 (5) (2009) 469–483.
- [2] T. Osa, J. Pajarinen, G. Neumann, J.A. Bagnell, P. Abbeel, J. Peters, et al., An algorithmic perspective on imitation learning, *Foundations and Trends® in Robotics* 7 (1–2) (2018) 1–179.
- [3] S. Schaal, Learning from demonstration, *Adv Neural Inf Process Syst* 9 (1996).
- [4] Y. Pan, C.-A. Cheng, K. Saigol, K. Lee, X. Yan, E.A. Theodorou, B. Boots, Imitation learning for agile autonomous driving, *Int J Rob Res* 39 (2–3) (2020) 286–302.
- [5] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, P. Abbeel, Deep imitation learning for complex manipulation tasks from virtual reality teleoperation, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 5628–5635.
- [6] D. Jiang, G. Li, Y. Sun, J. Hu, J. Yun, Y. Liu, Manipulator grabbing position detection with information fusion of color image and depth image using deep learning, *J Ambient Intell Humaniz Comput* 12 (12) (2021) 10809–10822.
- [7] X. Zhang, J. Liu, J. Feng, Y. Liu, Z. Ju, Effective capture of nongrasable objects for space robots using geometric cage pairs, *IEEE/ASME Trans. Mechatron.* 25 (1) (2019) 95–107.
- [8] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, G. Brain, Time-contrastive networks: Self-supervised learning from video, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 1134–1141.
- [9] F. Torabi, G. Warnell, P. Stone, Recent advances in imitation learning from observation, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 6325–6331, doi:10.24963/ijcai.2019/882.
- [10] P. Sharma, D. Pathak, A. Gupta, Third-person visual imitation learning via decoupled hierarchical controller, *Adv Neural Inf Process Syst* 32 (2019).
- [11] F. Torabi, G. Warnell, P. Stone, Behavioral cloning from observation, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, in: IJCAI'18, AAAI Press, 2018, pp. 4950–4957.
- [12] Y. Liu, A. Gupta, P. Abbeel, S. Levine, Imitation from observation: Learning to imitate behaviors from raw video via context translation, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 1118–1125.
- [13] W. Sun, A. Vemula, B. Boots, D. Bagnell, Provably efficient imitation learning from observation alone, in: International Conference on Machine Learning, PMLR, 2019, pp. 6036–6045.
- [14] T. Pierrot, N. Perrin-Gilbert, O. Sigaud, First-order and second-order variants of the gradient descent in a unified framework, in: I. Farkas, P. Masulli, S. Otte, S. Wermter (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2021*, Springer International Publishing, Cham, 2021, pp. 197–208.
- [15] T.M. Cover, J.A. Thomas, Elements of Information Theory (Wiley Series in Telecommunications and Signal processing), Wiley-Interscience, USA, 2006.
- [16] C.M. Bishop, N.M. Nasrabadi, Pattern recognition and machine learning, volume 4, Springer, 2006.
- [17] K.P. Murphy, Probabilistic Machine Learning: An Introduction, MIT Press, 2022.
- [18] S. Särkkä, Bayesian Filtering and Smoothing, Cambridge University Press, 2013.
- [19] K.P. Murphy, Probabilistic Machine Learning: Advanced Topics, MIT Press, 2023.
- [20] S. Ross, D. Bagnell, Efficient reductions for imitation learning, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 661–668.

- [21] S.K.S. Ghasemipour, R. Zemel, S. Gu, A divergence minimization perspective on imitation learning methods, in: Conference on Robot Learning, PMLR, 2020, pp. 1259–1277.
- [22] G.-H. Kim, J. Lee, Y. Jang, H. Yang, K.-E. Kim, Lobsdice: offline imitation learning from observation via stationary distribution correction estimation, arXiv preprint arXiv:2202.13536 (2022).
- [23] A. Edwards, H. Sahni, Y. Schroecker, C. Isbell, Imitating latent policies from observation, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, volume 97, PMLR, 2019, pp. 1755–1763.
- [24] Z. Zhu, K. Lin, B. Dai, J. Zhou, Off-policy imitation learning from observations, Adv Neural Inf Process Syst 33 (2020) 12402–12413.
- [25] C. Yang, X. Ma, W. Huang, F. Sun, H. Liu, J. Huang, C. Gan, Imitation learning from observations by minimizing inverse dynamics disagreement, Adv Neural Inf Process Syst 32 (2019).
- [26] F. Liu, Z. Ling, T. Mu, H. Su, State alignment-based imitation learning, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, 2020. <https://www.openreview.net/forum?id=rylrdxHFDr>
- [27] J. Ho, S. Ermon, Generative adversarial imitation learning, Adv Neural Inf Process Syst 29 (2016).
- [28] B. Kim, J. Pineau, Maximum mean discrepancy imitation learning, Robotics: Science and systems, 2013.
- [29] A. Giusti, J. Guzzi, D.C. Cireşan, F.-L. He, J.P. Rodríguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. Di Caro, et al., A machine learning approach to visual perception of forest trails for mobile robots, IEEE Rob. Autom. Lett. 1 (2) (2015) 661–667.
- [30] D. Jacobson, Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games, IEEE Trans Automat Contr 18 (2) (1973) 124–131.