Learning Search Strategies from Human Demonstrations

Dissertation (2014)

Submitted to the School of Engineering, Doctoral Program on Manufacturing Systems and Robotics

École Polytechnique Fédérale de Lausanne (EPFL)

in partial fulfillment of the requirements for the degree of Doctor of Philosophy

by

Guillaume de Chambrier

Thesis Committee:

Prof. Alireza Karimi, president of the jury
Prof. Aude Billard, thesis advisor
Prof. Hannes Bleuler, examiner
Prof. Jochen Steil, examiner
Prof. Ron Alterovitz, examiner

Lausanne, Switzerland
October, 2016

Chapter 1

Introduction

1.1 Motivation

Taking long term decisions or spontaneous reactive actions when presented with incomplete information or partial knowledge is paramount to the survival of any biological entity. Reasoning given uncertainty is a continuously occurring event throughout our livelihood. When considering long term decisions an abundance of examples come to mind; In economic investments uncertainty is, to the best of efforts, quantified and minimised. Reactive actions are just as common; When looking for the snooze button of an alarm clock, early in the morning, our hand seems to autonomously search the surrounding space, picking up sensory cues, gradually acquiring information which we utilise (or not) to guide us towards the button; Trying to connect a plug to a an occluded power socket under a desk, whilst being crouched, requires the integration of perceptions into a belief such to quantify the uncertainty which we can act upon to achieve the connection. Abilities close to these are not yet present in Artificial intelligence (AI) & robotics.

It is not yet fully understood how decisions are taken; yet alone under uncertainty. The difficulty is that two processes responsible for the synthesis of our actions, our beliefs and desires, are not directly measurable. The first attempt at modelling the humans decision making process was in mathematics & economics (Bernoulli (1954),Von Neumann and Morgenstern (1990)), where emphasis was on predicting discrete choices formulated as a gamble. It is only recently in Motion and Neuroscience that more incites have been gained.

Artificial intelligence & robotics considered early on uncertainty in decision making, where the predominant application domain was spatial navigation (Cassandra et al. (1996)). The problem is composed into of two parts: the construction and representation of a world model (the map) and a planner which can reason with respect to this model such to accomplish an objective. The world construction problem has attracted a large amount of research with many successfully applications in a wide spectrum of robotic domains (AUV,UAV,etc..). The planning problem is less well developed and is based on either representing the decision problem as a partially observable markov decision process (POMDP) which are notoriously difficult to solve for large scale problems, or through search

heuristics. The mapping problem can generally be solved when assuming the uncertainty is Gaussian and thus quantifiable by a few parameters and the uncertainty originates from the imprecision of the sensors. As for the planning problem solutions are feasible under the restrictive assumption of a discretization of the world, observations and actions of the robot. As a result there are very few examples where uncertainty is considered in an optimal decision make process when considering a continuous state, action and observation space.

In summary there are still open problems in decision making when considering partial observability, whilst the mapping problem has been studied under a constraining set of assumptions. In this thesis we address both problems under extreme levels of uncertainty. For the decision making side we leverage humans foresight and reasoning in a Learning from Demonstration (LfD) (Billard et al. (2008)) framework, which is used to transfer skills from an expert teacher (usually a human) to a robot. Examples include the transfer of kinematic task constraints, stiffness and impedance constraints and motion primitives, just to name a few. It has been shown, for the moment being, both humans and animals are far better at navigation than robots especially when uncertainty is present (Stankiewicz et al. (2006)). For the mapping problem we develop a Bayesian filter which is non-parametric and has no explicit representation of a joint distribution.

1.2 Contribution

In this thesis we bring to light two main ideas. The first is the transfer of human behaviour to robots in tasks where a lot of uncertainty in present, making them difficult to solve using traditional techniques. The second is a non-parametric Bayesian state space filter.

Throughout the work in this thesis we consider case studies in which vision is not available; leaving tactile and haptic information. This choice was made to induce a high level of uncertainty making it easier to study. As a consequence the tasks we consider are by nature, haptic and tactile searches.

1.2.1 Learning to reason with uncertainty as humans

A Markov Decision Process (MDP) allows to formulate a decision problem in terms of states, actions, a discount factor and a cost function. Given this formulation and a suitable optimisation method (dynamic programming, temporal difference, etc..) a set of optimal decision rules are returned, known as a policy. The benefit of this approach is that the policy is non-myopic and realises the importance of initial sub-optimal actions which might at first be necessary to achieve the task in the long run. A Partially Observable Markov Decision Process (POMDP), is a generalisation of an MDP to a hidden state space and

only observation are available relating to the state space. An exact solution to a POMDP is only feasible in simple toy problems (Thrun et al. (2005b)) and existing approximate solutions are tailored for discretized representation of states, actions and observations.

In this thesis we propose a Learning from Demonstration approach to solving the POMDP problem in haptic and tactile search tasks. Our hypothesis is that if we know the mental state of the human expert in terms of his believed location and observe his actions we can learn a statistical policy which mimics his behaviour. Since the human's beliefs are not directly observable we infer them by assuming that the way we integrate behaviour is similar to a Bayesian filter. There is evidence both in cognitive and neuroscience that this is the case (Bake et al. (2011)). From the expert human demonstrations of the task we learn a cognitive model of the humans decision process by learning a generative joint distribution over his beliefs and actions. The generative distribution is then used as a control policy. By this approach we are able to have a policy which can handle uncertainty similarly to humans.

1.2.2 Non-parametric Bayesian state space filter

Simultaneous Localisation and Mapping (SLAM) is concerned with the development of filters to accurately and efficiently infer the state parameters (position, orientation,...) of an agent and aspects of its environment, commonly referred to as the map. It is necessary for the agent to achieve situatedness which is a precondition to planning and reasoning. The predominant usage of SLAM algorithm make the assumption that uncertainty is related to the noise in the sensor measurements. In our haptic search tasks there is no visual information and a very large amount of uncertainty. Most of the sensory feedback is negative information, a term used to denote the non event of a sensor response from the objects (aka landmarks) in question. In the absence of recurrent sightings or direct measurements of objects there are no correlations from the measurement errors which can be exploited.

In this thesis we propose a new SLAM filter, which we name Measurement Likelihood Memory Filter (MLMF), in which no assumptions are taken with respect to the shape of the uncertainty (it can be Gaussian, multi-modal, uniform, etc..) and motion noise. From the loose assumptions we stipulate regarding the marginals, we adopt a histogram parametrisation (this is considered non-parametric because a change in a parameter has a local effect). The conceptual difference between the MLMF and standard SLAM filters such as EKF is that we avoid representing the joint distribution since it would entail a shattering space and time complexity. This is achieved by keeping track of the history of measurement likelihood functions. We demonstrate that our approach gives the same filtered marginals as a histogram filter. In such a way we achieve a Bayes

filter which has both linear space and time complexity. This filter is well suited to tasks where the landmarks are not directly observable.

1.2.3 Reinforcement learning in belief space

We propose a Reinforcement Learning framework for the task of searching and connection a power plug to a socket, with only haptic and tactile information. We previously addressed this setup by learning a generative model of the beliefs and actions with data provide by human demonstrations following the LfD approach. However, it is usually the requirement in such setups that the teach is an expert, with few notable exceptions (Rai et al. (2013)). Since we were solely learning a statistical controller, bad and good demonstrations will be mixed in together. By introducing a cost function representing the task we can explicitly have a quality metric of the provided demonstrations. In this way we can optimise the parameters of our generative model to maximise the cost function. In this LfD Reinforcement Learning setup with a very simple cost function we can have a significant improvement of our a policy.

1.3 Thesis outline

The thesis is structured accordingly to the three main contributions outlined in the previous section, and three will have their individual chapter. We first provide and background chapter situating our work in the scientific community and give a conclude with a discussion of the contributions and impact of our work.

In this chapter we review the background literature which are the pillars of this thesis, namely: *Decision Theory*, *Theory of Mind* and *Reasoning under uncertainty*. These three topics are the root nodes of their own respective fields and we do not seek to do all of them justice individually, but highlight their relevance and contribution to our work.

Chapter 2

Background

Planning and reasoning under uncertainty is central to robotic and artificial intelligence research and has been an active area of research for decades. It is an umbrella term which touches a wide spectrum of fields: *economics*, *psychology*, *cognitive science*, *neuroscience*, *robotics* and *artificial intelligence*. The work in this thesis relies on results and assumptions made in cognitive and neuroscience with respect to our beliefs and how we act given them. We complement these results by introducing them in a new light to the field of robotics and demonstrate how the human reasoning and belief system can be used in situations where the state space is partially observable. The second main theme our work builds on is state space estimation. The third component acting given uncertainty in robotics. We make use of results from all three fields. We provide a background overview of acting under uncertainty and situate our work within the state of the art.

This chapter unfolds as follows:

2.1 Decisions under Uncertainty

In this section we introduce and frame the problem we seek to solve in generic terms. We are concerned with finding a sequence of actions which will lead to the successful outcome of a problem being considered; this is the most generic definition.

There are two key attributes which can make this problem difficult: stochastic actions and latent states. Stochastic actions, when applied in the same state will not always result int the same outcome. This type of uncertainty can arise from many sources; the outcome of chaotic actions are impossible to predict with certainty, think of throwing a die or flipping a coin; In outdoor robotics the terrain might lead to slippage, causing the robot to skid or underwater currents might drastically offset the position of an UAV; In articulated robots the friction between joints can accumulate to a large error in the end-effector position (especially true for cable driven robots). The second source of uncertainty is when the underlying state is partially known, in the sense that we do not have all the necessary information to reliably determine the state beyond reasonable doubt. In robotics this uncertainty can arise from inadequate or noisy sensors.

If the environmental conditions in which the robot is located is humid, misty or dark. It can make it difficult for the robot to ascertain its position and to plan how to achieve a given objective.

The uncertainty of the state and actions have to be quantified. The predominant approach is to represent them by probabilities. For instance the application of a forward action (for a wheeled robot) will result in a new position further ahead and a position to the right (due to slippage) with some probability. An observation through the robots sensors will result in probability distribution over the robots probable location. Given this quantification of action and observation uncertainty in terms of a probability distribution over the state, the agent must now take actions towards accomplishing its goal. To take a decision the agent must assign a utility to the outcome of his actions. The utility is to indicate a preference over the outcomes and when combined with probabilities leads to decision theory.

2.1.1 Decision theory

The central question of decision theory is; *how do people take decisions when faced with uncertain outcomes ?* Interest in such questions were typically centred around economical questions such as deciding what should be an appropriate investment or wager for a particular gamble. It was noted that the expected monitory outcome of a gamble as a mean of basing a decision, would often lead to a course of action which contradicts common sense. A famous example is the St. Petersburg paradox. Daniel Bernoulli proposed a solution to the problem by introducing the notion of a *utility* function in which he claimed that people should base their decision on the expected utility instead of solely the monetary outcomes of the gamble.

> "...the value of an item must not be based on its price, but rather on the utility it yields."
>
> — Daniel Bernoulli,Bernoulli (1954)

The introduction of a utility function takes into account that the net worth of a person will influence their decision since they weigh the gain differently. The utility function introduced by Bernoulli was the logarithm of the monetary outcome $x \in \{5\$, 10\$, 25\$\}$ weighted by their probability $p(x)$ which results in an expected utility, Equation 2.1.

$$U(x) = \mathop{\mathbb{E}}_{x \sim p(x)} \{u(x)\} = \sum_{x \in X} p(x) \underbrace{\log(x)}_{u(x)} \tag{2.1}$$

Different utility functions characterise different levels of risk. When the it is concave as it for Bernoulli's utility function the person will be **risk-averse**.

Risk-averse means that a gambler would prefer the utility of a sure outcome instead of taking a gamble who's expected utility is the same as the one of the sure outcome. This was the first introduction of a utility function. It is later in 1944 that von Neumann and Morgenstern (Von Neumann and Morgenstern (1990)) axiomised Bernoulli's utility function and proved that if a decision maker has a preference over a set of lotteries[1] which satisfy four axioms (completeness, transitivity, continuity, independence) then there exists a utility function who's expectation preserves this preference. An agent whose decisions can be shown to maximise the vNM expected utility are said to be **rational** and otherwise **irrational**. This is the theoretical basis of most economic theory, it is a **normative** model of how people should behave given uncertainty. A drawback to the vNM theory of rationality is that it lacks the descriptive power to model peoples actual decisions. It predicts that when we are presented with a choice with two different lotteries we will chose the lottery which yields the maximum expected utility. There notable human studies ([citation]) which consistently demonstrate that we do not always act as a vNM agent. Decision Theory models which try to describe how we behave and not how we should behave are known as prescriptive models, (see (**??**) and the prospect model (**?**)).

2.1.2 Beliefs & desires

2.2 Partially Observable Markov Decision Process

A POMDP is a popular approach for formulating a decision making process under both motion and measurement uncertainty; In Figure 2.1 we describe all of the components necessary for a POMDP.

Since the states are not observable, the agent cannot choose its actions based on the state. The explicit representation of the past events is typically memory expensive. Instead it is possible to summarize all relevant information from previous actions and observations in a probability distribution over the state space, known as the belief state.

Because the state space is partially observable the expected reward has to be computed for each possible history of states, actions and observations. All approaches in the literature instead encapsulate all these possible histories into a belief state $b_t(x_t)$ which is a probability distribution (referred to in the POMDP literature as an information state, I-state) over the state space x_t and use this new state description to cast the POMDP into a *belief*-MDP (states are probability distributions, beliefs).

[1] the term lottery refers to a probability distribution in the original text.

POMDP & *belief*-MDP

States, Actions, Observations

Transition function: $\mathbf{p(x_{t+1}|x_t, u_t)}$
The state transition function models the uncertainty originating from motion noise and is represented by the conditional probability distribution (or likelihood) function, $p(x_{t+1}|x_t, u_t) \in \mathbb{R}$, which gives the probability of moving to state x_{t+1} given that action u_t was applied in state x_t.

Observation function: **$\mathbf{p(y_t|x_t)}$**
The observation function returns the probability or likelihood of the current observing y_t given a known state x_t. It is modelled by the conditional distribution $p(y_t|x_t) \in \mathbb{R}$.

Belief: $\mathbf{b_t(x)}$
A belief is probability distributions, $b_t(x)$, over the state space X and quantifies both motion and observation uncertainty.

State space estimator: $\mathbf{b_t(x) = \tau(b_{t-1}(x), u_{t-1}, y_t)}$
Updates a belief given a motion and observation, it makes use of both the motion and observation functions defined in the POMDP. The state space estimation function, τ, can be any kind of state space filter such as an Extended Kalman Filter (EKF) or a Particle Filter (PF).

Reward function: $\mathbf{R(x_t, u_t)}$
The reward becomes a function of the belief $R(b_t, u_t)$ which is the expected value of the original reward function $\mathbb{E}_{x_t \sim b_t}[R(x_t, u_t)]$.

Discount factor $\gamma \in [0, 1]$;

Figure 2.1: Description of the individual elements necessary to formalise a POMDP and a *belief*-MDP.

The new description of the problem is now in terms of belief space $< \mathcal{B}, U, \tau, R, \gamma >$ where $\mathcal{B}$ is the set of all possible beliefs and The reward becomes a function of the belief $R(b_t, u_t)$ which is the expected value of the original reward function $\mathbb{E}_{x_t \sim b_t}[R(x_t, u_t)]$, The goal is to find an optimal action for each belief such that the policy $\pi(b_t, u_t)$ maximises the expected reward, Equation 2.2.

$$V^{\pi^*}(b_{t-1}) = \max_{u_{t-1}} \left[R(b_{t-1}, u_{t-1}) + \gamma \cdot \mathbb{E}_{y_t} \left[V^{\pi}(b_t) \right] \right] \tag{2.2}$$

From considering the decision belief tree of the POMDP, Figure **??**, we can appreciate the complexity of the problem of finding an optimal policy. Given a discrete set of actions and observations to update the belief b_1 we have to consider a time complexity of $\mathcal{O}(|U||Y|^T)$ where T is the depth of the tree (the planning horizon). Given that we have a finite set of belief the complexity solving the POMDP is $\mathcal{O}(|\mathcal{B}||U||Y|^T)$.

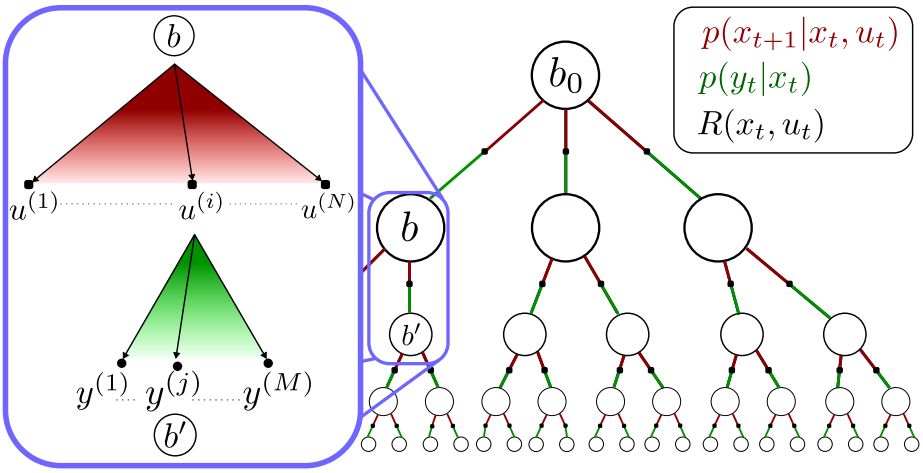

Figure 2.2: ad

2.3 State of the art

Hansen (1998)

Tree search

Planning

$b = (\mu, \Sigma)$ He et al. (2008),Prentice and Roy (2009)

Optimal control

$b = (\mu, \Sigma)$

Erez and Smart (2010), Brooks and Williams (2011), Platt et al. (2010)

Optimal control methods represent the belief by a Gaussian function

Martinez-Cantin et al. (2009),Spaan and Vlassis (2005),Thrun et al. (2005a)

Hauser (2010)

Ross et al. (2008)

He et al. (2011)

2.4 Summary

References

Chris Bake, Joshua. Tenenbaum, and Rebecca Saxe. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Thirty-Third Annual Conference of the Cognitive Science Society*, pages 2469–2474, 2011. 1.2.1

D. Bernoulli. Exposition of a New Theory on the Measurement of Risk (1748). *Econometrica*, 22(1):23–36, 1954. 1.1, 2.1.1

A. Billard, S. Calinon, R. Dillmann, and S. Schaal. Robot programming by demonstration. In B. Siciliano and O. Khatib, editors, *Handbook of Robotics*, pages 1371–1394. Springer, Secaucus, NJ, USA, 2008. 1.1

Alex Brooks and Stefan Williams. A monte carlo update for parametric pomdps. In Makoto Kaneko and Yoshihiko Nakamura, editors, *Robotics Research*, volume 66 of *Springer Tracts in Advanced Robotics*, pages 213–223. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-14742-5. doi: 10.1007/978-3-642-14743-2_19. URL http://dx.doi.org/10.1007/978-3-642-14743-2_19. 2.3

A. R. Cassandra, L. P. Kaelbling, and J. A. Kurien. Acting under uncertainty: discrete bayesian models for mobile-robot navigation. In *Intelligent Robots and Systems '96, IROS 96, Proceedings of the 1996 IEEE/RSJ International Conference on*, volume 2, pages 963–972 vol.2, Nov 1996. doi: 10.1109/IROS.1996.571080. 1.1

Tom Erez and William D. Smart. A scalable method for solving high-dimensional continuous pomdps using local approximation. In *Conf. on Uncertainty in Artificial Intelligence*, 2010. 2.3

Eric A. Hansen. Solving pomdps by searching in policy space. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pages 211–219, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-555-X. URL http://dl.acm.org/citation.cfm?id=2074094.2074119. 2.3

Kris Hauser. Randomized belief-space replanning in partially-observable continuous spaces. In David Hsu, Volkan Isler, Jean-Claude Latombe, and Ming C. Lin, editors, *WAFR*, volume 68 of *Springer Tracts in Advanced Robotics*, pages 193–209. Springer, 2010. 2.3

Ruijie He, S. Prentice, and N. Roy. Planning in information space for a quadrotor helicopter in a gps-denied environment. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 1814–1820, May 2008. doi: 10.1109/ROBOT.2008.4543471. 2.3

Ruijie He, Emma Brunskill, and Nicholas Roy. Efficient planning under uncertainty with macro-actions. *J. Artif. Int. Res.*, 40(1):523–570, January 2011. ISSN 1076-9757. URL http://dl.acm.org/citation.cfm?id=2016945.2016959. 2.3

Ruben Martinez-Cantin, Nando de Freitas, Eric Brochu, JosÃl' Castellanos, and Arnaud Doucet. A bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot. *Autonomous Robots*, 27(2):93–103, 2009. ISSN 0929-5593. doi: 10.1007/s10514-009-9130-2. URL http://dx.doi.org/10.1007/s10514-009-9130-2. 2.3

R. Platt, R. Tedrake, L. Kaelbling, and T. Lozano-Perez. Belief space planning assuming maximum likelihood observations. In *Proceedings of Robotics: Science and Systems*, Zaragoza, Spain, June 2010. 2.3

S. Prentice and N. Roy. The belief roadmap: Efficient planning in belief space by factoring the covariance. *International Journal of Robotics Research*, 8 (11-12):1448–1465, December 2009. 2.3

Akshara Rai, Guillaume De Chambrier, and Aude Billard. Learning from failed demonstrations in unreliable systems. In *Humanoid Robots (Humanoids), 2013 13th IEEE-RAS International Conference on*, pages 410–416. IEEE, 2013. 1.2.3

StÃl'phane Ross, Joelle Pineau, SÃl'bastien Paquet, and Brahim Chaib-draa. Online planning algorithms for pomdps. *Journal of Artificial Intelligence Research*, 2008. 2.3

Matthijs T. J. Spaan and Nikos Vlassis. Planning with continuous actions in partially observable environments. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3469–3474, Barcelona, Spain, 2005. 2.3

B.J. Stankiewicz, G.E. Legge, J.S. Mansfield, and E.J. Schlicht. Lost in virtual space: Studies in human and ideal spatial navigation. *Journal of Experimental Psychology: Human Perception and Performance.(under review)*, 32(3):688–704, 2006. 1.1

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005a. ISBN 0262201623. 2.3

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005b. 1.2.1

John Von Neumann and O. Morgenstern. *The theory of games and economic behavior*. Princeton, 3 edition, 1990. 1.1, 2.1.1