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 (?). 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 (?) 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 (?). 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 we take decisions when faced with uncertain outcomes ?* To answer such a question we need to ground the attributes which are involved when we take a decision, namely our **beliefs** and **desires**. Beliefs reflect a degree of knowledge we have about the world in which the degree is ascertained by the amount of evidence we have in support of our beliefs. Epistemology studies in great detail the relationship between truth, beliefs and knowledge. We will not go into a philosophical discussion of their interplay, but make use of the following; if we have sufficient evidence in support of our beliefs and they represent the truth then we consider them to be a **rational belief**. As for desires they are linked to our disposition to take action to achieve them; for example if I want to switch of my alarm clock I have to look for it in the last area I believed it to be. These two attributes, beliefs and desires, are used to frame a decision problem. Early work in decision theory assumed that the problem was well grounded and focused on finding what are the **rational choices** to take given our beliefs to achieve our desires.

Early interest in such questions were typically centred around economics 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. In this paradox a bookmaker proposes you the following gamble. An initial pot starts with a content of 2£, the bookmaker proceeds to flip a fair coin until the first appearance of a tails which ends the game. Until the occurrence of the first tails the money in the pot doubles after every toss. Once the game ends you leave with the content of the pot. As an avid gambler and expected value max-

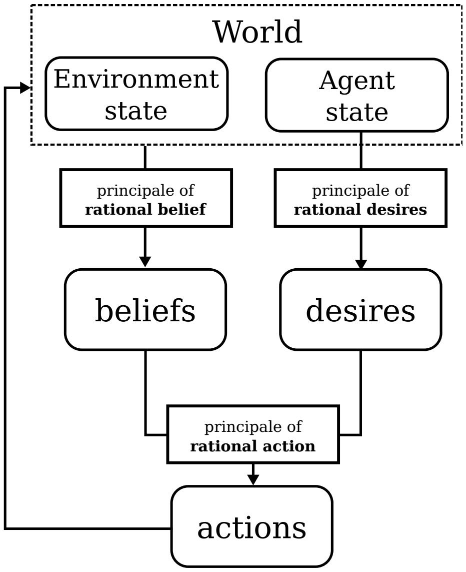

World
Environment state
Agent state
principale of rational belief
principale of rational desires
beliefs
desires
principale of rational action
actions

Figure 2.1: ad

imiser how much would you be willing to pay to enter this gamble ? A value maximiser would computed the expected monetary outcome. The amount of money increases by 2^n£, where n is the number of non-final tosses and and the probability of reaching n is $1/2^n$. In this case the expected monitory outcome is an infinite number,

$$\mathop{\mathbb{E}}_{p(n)}\{£\} = \underbrace{\frac{1}{2}2£}_{\text{first toss}} + \frac{1}{4}4£ + \cdots = \sum_{n=1}^{\infty} \frac{2^n}{2^n}£ = \infty£$$

So your expected gain or return for paying to enter such game is an infinite amount of money, so in principal if you were seeking to maximise your expected return value you would be willing to pay an amount close to infinity. This does not seem a good decision rule; no person in the world would be willing to pay more than 1£ to enter such game.

Nicola Bernoulli proposed a solution to the problem (later published by his brother Daniel (Bernoulli (1954))) by introducing the notion of a **utility function**, and he claimed that people should base their decision on the expected utility instead of solely the monetary outcomes of a gamble.

> “...the value of an item must not be based on its price, but rather on the utility it yields.”
>
> — Daniel Bernoulli

The introduction of a utility function takes into account that the net worth of a person will influence their decision since different people (in terms of their monetary worth) will weigh the gain differently. The utility function introduced by Bernoulli was the logarithm of the monetary outcome $x \in X$ weighted by their probability $p(x)$ which results in an expected utility,

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

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**, when linear **risk-neural** and convex **risk-seeking**. 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. It is also the basis of most if not all decision making, cogitative architectures and control policies in AI and robotics (to the best of the authors knowledge).

An aspect to keep in mind regarding the vNM model is that it is normative; it states what should be a rational decision or behaviour. As a result it is not always consistent with human behaviour. There is great debate regarding the predictions made by vNM models with respect to ours. There have been many studies both demonstrating divergence between the models predictions and our observed behaviour but also supporting evidence that it does reflect the output of our decision making process. Reasons for divergence have been attributed to the way we weigh probabilities and how the decision problem is framed. But probably the most important aspect is that in most decisions we are faced with the quantification and rationality of our beliefs might not be adequate and limitations of our working memory will come into play in the final decision.

Nevertheless vNM agents are predominantly used in AI and robotics as a means of implementing a decision making process or a control policy. In psychology and cogitative science vNM agents are a used for comparing human behaviour against an optimal strategy (by optimal we mean it is rational in the vNM sense).

2.1.2 Beliefs & desires

2.2 Sequential decision process

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

Notation	**Definitions**
$x_t \in \mathbb{R}^3$	Cartesian state space position of the agent.
$y_t \in \mathbb{R}^M$	Observation/measurement from the agents sensors.
$a_t \in \mathbb{R}^3$	Action, usually the Cartesian velocity of the end-effector of the agent.
X, Y, A	State, observation and action random variables where x, y and a are realisation.
$p(x_t)$	Short hand notation for a probability density function, $p_X(x_t)$.
$x_{0:t}$	$\{x_0, x_1, \cdots, x_{t-1}, x_t\}$, history up to time t.
$p(x_t \| y_{0:t}, a_{0:t})$	Filtered probability distribution over the state space given the action and observation history.
$b_t \in \mathbb{R}^L$	Belief state, a function of the filtered distribution $b(p(x_t \| y_{0:t}, a_{0:t}))$ which will be written as b_t for simplicity.
$\pi_{\boldsymbol{\theta}}(a_t \| \cdot)$	Probabilistic policy, $a_t \sim \pi_{\boldsymbol{\theta}}(a_t \| \cdot)$
$r(x) \in \mathbb{R}$	Reward function, returns the utility of being in state x. It can also be dependent on the action, $r(x, a)$.
$\gamma \in [0, 1)$	Discount factor, the closer to one the more later utilities/rewards are considered. When set to zero, only immediate rewards are considered which would result in a myopic greedy agent.
$p(x_{t+1} \| x_t, a_t)$	State transition function, returns the likelihood/probability of reaching state x_{t+1} given that action a_t is applied in state x_t.
$p(y_t \| x_t)$	Observation/measurement model, returns the likelihood/probability of observing y_t given that the agent is in state x_t.
$\tau(b_{t-1}(x), u_{t-1}, y_t)$	Updates a belief given a motion and observation, it makes use of both the motion and observation functions. 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).

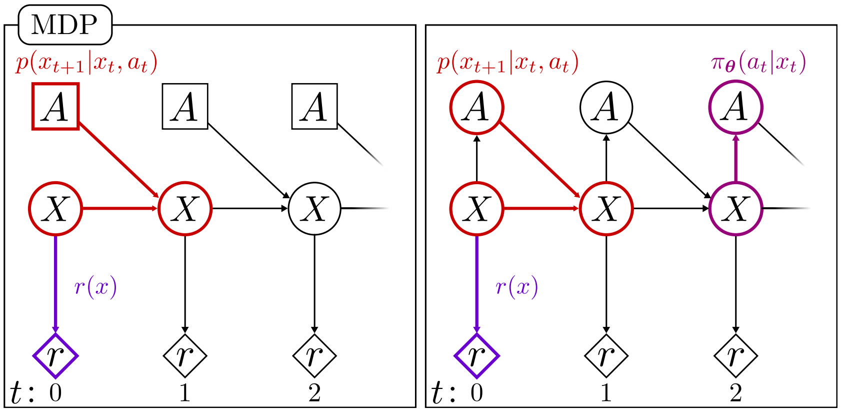
Table 2.1: Definition of common variables used.

When referring outright to decision theory with no extensions, we usually are talking about a one-shot non-temporal decision. However many interesting decision problems are sequential. In such a situation we must consider the effect current decisions will have on future decisions. Expected utility theory (part of decision theory) is extendible to a temporal decision problem. There are however a two subtle but important differences between the temporal and non-temporal decision problems. The first is the utility, in the one time step problem an outcome has one utility assigned to it, $u(x)$. Now a utility has to be assigned to a sequence of outcomes, $u(x_{0:T})$, where T is the number of sequential decisions taken. The utility of a sequence is the sum of the individual outcomes themselves. However if the decision problem is non terminating this will lead to an unbounded utility. To bound the utility a discount factor $\gamma \in [0, 1)$ is introduced and the new utility function becomes:

$$u(x_{0:T}) = \sum_{t=0}^{T} \gamma^t u(x_t) \tag{2.1}$$

The discount factor allows to control the importance later utilities have on the final utility. If the discount factor is set to zero we recover the original one-shot utility function and if we were to take actions which maximised the expected utility we would not be considering at all the effect later decisions have on our overall utility. An agent reasoning in such a way is called myopic. The second is the way in which probabilities are assigned to outcomes, this was $p(x)$ in the decision theory utility function formulation. Now because of the sequential nature of the problem we consider a conditional state transfer probability distribution $p(x_{t+1}|x_t, a_t)$ which models the probability of going from state x_t to x_{t+1} given that action a_t is taken. This particular representation of a sequential decision problem is called a **Markov Decision Process (MDP)** and to be more exact a first order MDP. The necessary models are the state transition and utility functions. The assumption of such a model is that all necessary information to take a decision is encoded in the current state and there is no need to consider the history of state transitions when taking a current decision. In Figure 2.2 we illustrate two graphical representations of a MDP, which are known as **Dynamic Bayesian Networks (DBN)**. A DBN represents the the temporal relationship and conditional dependence between random variables, decisions and utilities, which are represented by circles, squares and diamonds. For the MDP to the left the actions are not stochastic, whilst for the MDP on the right the actions taken are governed by a stochastic **policy**, $\pi_{\boldsymbol{\theta}}(a_t|x_t)$. A policy represents the decision process of an agent, given a state it will output an action. A stochastic policy means that given the same input they will produce different outputs. A policy is considered optimal when it maximises the expected utility function, it is optimal in the vNM sense.

Solving a MDP means finding a policy whose actions in any given state will always maximise the expected utility. Such a policy is usually denoted as π^*,



(a) off-policy **(b)** on-policy

Figure 2.2: Dynamical Bayesian Network of a Markov Decision Process; it encodes the temporal relation between the random variables (circles), utilities (diamond) and decisions (squares). The arrows specify conditional distributions. In **(a)** the decision nodes are not considered random variables whilst in **(b)** they are. From these two DBN we can read off two conditional distributions, the state transition distribution (in red) and the action distribution (in purple).

the optimal policy. As in decision theory, the expected utility is the utility of a sequence of states $u(x_{0:T})$ weighted by its probability . The graphical representation allows to read off directly the probability of a sequence of state transitions and actions $(x_{0:T}, a_{0:T-1})$, Equation 2.2.

$$p(x_{0:T}, a_{0:T-1}) = p(x_0) \prod_{t=0}^{T-1} p(x_{t+1}|x_t, a_t) \tag{2.2}$$

$$u(x_{0:T}) = r(x_0) + \gamma r(x_1) + \cdots + \gamma^{T-1} r(x_{T-1}) + \gamma^T r(x_T) \tag{2.3}$$

We are interested in finding the sequence of action, $a_{0:T}$, which will maximise the expected utility function:

$$\underset{a_{0:T-1}}{\operatorname{argmax}} U(x_{0:T}, a_{0:T-1}) = \max_{a_0} \sum_{x_1} \cdots \max_{a_{T-1}} \sum_{x_T} \Big(p(x_{0:T}, a_{0:T-1}) u(x_{0:T}) \Big) \tag{2.4}$$

Solving the above directly in its current form would have an exponential complexity. By making use of the first order markov assumption and that current rewards do not dependent future rewards, the sums in Equation can be rearranged and and recursive patter will emerge which we can take advantage off. Lets start from the last time step and move out of the sum all elements which do not depend on T and $T-1$, which results in Equation 2.5

$$\underset{a_{0:T-1}}{\operatorname{argmax}} U(x_{0:T}, a_{0:T-1}) = \max_{a_0} \sum_{x_1} \cdots \max_{a_{T-2}} \sum_{x_{T-1}} p(x_{0:T-1}, a_{0:T-2})$$
$$\left(u(x_{0:T-2}) + \gamma^{T-1} \left(r(x_{T-1}) + \gamma \max_{a_{T-1}} \sum_{x_T} p(x_T|x_{T-1}, a_{T-1}) r(x_T) \right) \right) \tag{2.5}$$

From the rearrangement we notice that Equation 2.5 has the same functional form as Equation 2.4, except that the recursive component can be summarised by Equation 2.6, which is known as the **Bellman** optimal equation.

$$V^*(x_t) := r(x_t) + \gamma \max_{a_t} \sum_{x_{t+1}} p(x_{t+1}|x_t, a_t) V(x_{t+1}) \tag{2.6}$$

For the terminal state $V_T(x_T) = r(x_T)$. The bellman equation is a means of solving a sequential decision problem through use of dynamic programming. It says the the utility of the current state is based on the immediate reward and the discounted maximum utility of the next state. Making use of this recursion reduced the computation complexity is quadratic in the number of states, $\mathcal{O}(T\,|A|\,|X|^2)$. To find the optimal value and subsequent policy an approach would be to repeatedly apply the bellman equation to each state until the value function converges. What makes the problem hard to solve is maximisation over the actions. This induces two problems, the first is that the optimisation is nonlinear and the second is that if the action space is continuous the maximisation will be expensive to compute. This brings use to the two main approaches to solving a the MDP process: **off-policy** and **on-policy**. Off-policy methods solve directly for the optimal value function $V^*(x)$ and perform the maximisation over the actions, **Value-Iteration (VI)** is such a method. On-policy approaches find the optimal value and policy through repeating **policy evaluation** and **improvement** steps. In the policy evaluation the value or utility of a policy is found through solving the on-policy version of the Bellman equation:

$$V^\pi(x_t) := r(x_t) + \gamma \sum_{a_t} \pi_{\boldsymbol{\theta}}(a_t|x_t) \sum_{x_{t+1}} p(x_{t+1}|x_t, a_t) V(x_{t+1}) \tag{2.7}$$

In the policy improvement step the policy is made more greedy by maximising the value function. Through the repetition of these two steps both the value function and policy converge to the optimal. On-policy methods are preferred in settings where the action space is highly continuous, such as in robotics. Using dynamic programming is however not the method of choice since it requires multiple passes through the entire state space and for this it is necessary to have the model of the state transition a priori. Instead **Reinforcement Learning (RL)** methods are used to find an optimal value and policy. RL is a sample based approach in which an agent interacts with the environment gathering examples of state transitions and rewards (the utility) and uses them to gradually solve the bellman equation.

We introduced the formulation of a sequential decision process in the form of a MDP model and showed how an optimal policy and value function are obtained through maximising the expected utility. The re-arrangement of the sums via variable elimination allows to take advantage of a recursive structure present in the markov chain. The recursive component turns out to be the Bellman optimal equation, which when solved (via dynamic programming or reinforcement learning) results in an optimal value and policy function. A MDP models the uncertainty inherent in the state transition but not the uncertainty of the state. The MDP assumes that the state space is always fully observable, which is a strong assumption. In robotics the on bored sensors return an estimate of the state with a certain amount of uncertainty associated with it. To take this additional uncertainty into consideration the MDP has to accommodate it. This leads to a Partially Observable Markov Decision Process (POMDP).

2.2.1 POMDP

A POMDP is a popular approach for formulating a sequential decision process in which both motion and observation uncertainty is considered. In this partially observable setting the agent does not know with exactitude the state of the environment, but is able to observe it through his **sensors**. We mathematically define a sensor as being a function of the state space, Equation 2.8.

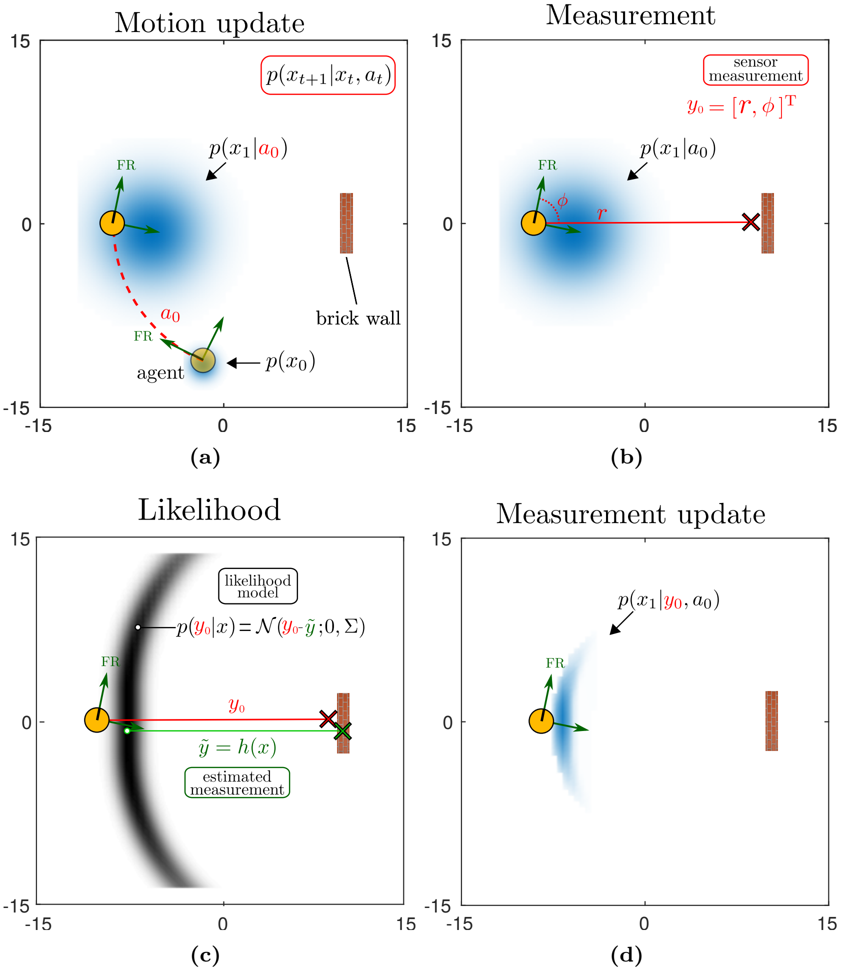
$$y_t = h(x_t) + \epsilon_t \tag{2.8}$$

The sensor function $h(\cdot)$ can be linear or non-linear and the additive noise term ϵ_t can be Gaussian (usually the case), non-Gaussian, state dependent or not. In this setting the state x is a latent hidden variable and we can only get information about it via the observations, y. The uncertainty of the latent state is quantified in a probability distribution over the state space, $p(x)$. This probability distribution represents all the hypothetical positions in the world in which the agent can be. In Figure 2.3 **(a)** an agent is located in a environment a in a yard containing a wall. Initially the agent is confident regarding his position; his state uncertainty $p(x_0)$ is low, represented by the blue probability density. However during a circular displacement the agent skids and the state uncertainty is increased by the state transition function, $p(x_{t+1}|x_t, a_t)$. To reduce the uncertainty, the agent takes a measurement, y, with his sensors which provide range, r, and bearing, ϕ, information of the wall, see Figure 2.3 **(b)**. The agent uses the model of his sensor, known a priori, to deduce all possible locations in the world, x, where the current measurement could have originated from. This model is known as the measurement likelihood function, Equation 2.9

$$p(y_t|x_t) = \mathcal{N}(y_t - h(x_t); 0, \Sigma) \tag{2.9}$$

The measurement likelihood function makes use of the measurement function $h(x)$ and it models the noise in the sensor. In this case the parameters of the noise model, ϵ_t, is Gaussian with mean zero and covariance Σ. Typically the parameters of the measurement likelihood function are learned a priori.

In Figure 2.3 **(c)** the likelihood is illustrated; the dark regions indicate areas



(a)

(b)

(c)

(d)

Figure 2.3: **(a)** An agent is located to the south west of a brick wall, it is equipped with a range sensor. The agent takes a forward action, but skids which results in a high increase of the uncertainty.**(b)** The agent takes a measurement, y_0, of this distance to the wall; because his sensor is noisy his estimate is off. **(c)** The agent uses with his measurement model to evaluate the plausibility of all locations in the world which would result in a similar measurement; illustrated by the likelihood function $p(y_0|x_0)$. **(d)** The likelihood is integrated into the probability density function; $p(x_0|y_0) \propto p(y_0|x)p(x_0)$.

Bayesian filter

The Bayesian filter turns a prior probability distribution over the state space, $p(x_t|y_{0:t-1}, a_{0:t-1})$, to a posterior $p(x_t|y_{0:t}, a_{0:t})$ by incorporating both motion and measurement. Applied recursively it keep a probability distribution over the state space which considers all the past history of actions and observations.We define the application of these two steps by the filter function τ, which returns takes the current belief, applied action and measurement to return the next belief, b_{t+1}.

Motion update

$$p(x_t|y_{0:t-1}, a_{0:t}) = \int p(x_t|x_{t-1}, a_{t-1})\, p(x_t|y_{0:t-1}, a_{0:t-1})\; da_{t-1} \tag{2.10}$$

Measurement update

$$p(x_t|y_{0:t}, a_{0:t}) = \frac{1}{p(y_t|y_{0:t-1}, a_{0:t})} p(y_t|x_t)\, p(x_t|y_{0:t-1}, a_{0:t}) \tag{2.11}$$

$$p(y_t|y_{0:t-1}, a_{0:t}) = \int p(y_t|x_t)\, p(x_t|y_{0:t-1}, a_{0:t}) dx_t \tag{2.12}$$

Filter function

$$b_{t+1} := \tau(b_t, a_t, y_t) \tag{2.13}$$

Figure 2.4: Bayesian state space filter.

of high likelihood, which are possible locations from which the sensor measurement could have originated from. The value of the measurement likelihood function is then integrated into the state space probability density function.

These two steps: **motion** and **measurement** updates are part of a state estimation process called a **Bayesian state space filter**, which we formalise below in Equation 2.10-2.11.

The motion model, Equation 2.10, updates the position of the probability distribution according to the applied action, a_t, and adds uncertainty by increasing the spread of the distribution. The measurement information is the incorporated by Equation 2.11. The measurement likelihood always decreases the uncertainty or leaves it constant. It never results in an increase of uncertainty. The Bayesian state space filter is such an important component to belief space decision making that we define it by the filter function, $\tau(b_t, a_t, y_t)$, which takes as input the current belief, applied action and sensed measurement and returns the resulting belief b_{t+1}. The state space filter is an essential component to a POMDP as it will become apparent.

With the latent state and its relation to the observation variable and the Bayesian filter defined, we can introduce the POMDP model in Figure 2.5 (*left*). It has the same markov chain structure present in MDP, introduced in the

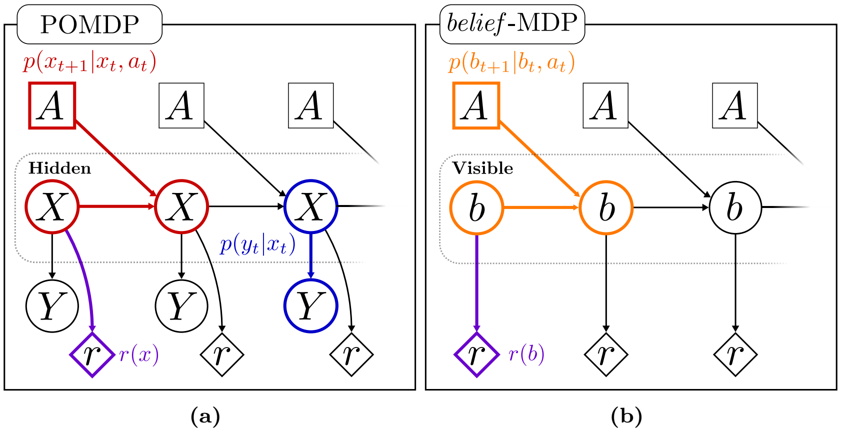


Figure 2.5: **(a)** POMDP graphical model. The state space, X, is hidden, but is still partially observable through a measurement, Y. **(b)** belief-MDP, the POMDP is cast into a belief Markov Decision Process. The state space is a probability distribution, $b(x_t) = p(x_t)$, (known as a belief state) and is no longer considered a latent state. The original state transition function $p(x_{t+1}|x_t, a_t)$ is replaced by a belief state transition, $p(b_{t+1}|b_t, a_t)$. The reward is now a function of the belief.

previous section, but the state space X is latent and a new layer of observation variables Y is present.

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 $b(x_t)$ (for short notation b_t) which is a probability distribution (also referred to 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). By casting the POMDP to a *belief*-MDP the state space is considered observable and we recover the same structure as in the standard MDP problem.

Because we are working with in a belief-space the reward function has to be adapted to:

$$r(b_t) = \int_{x_t} r(x_t)\, b(x_t)\; dx_t \tag{2.14}$$

which is the expected reward $r(b_t) = \mathbb{E}_{b_t}\{r(x_t)\}$. The goal as before is to find a sequence of actions which will maximise the expected utility. Since our *belief*-MDP has the same structural form as the MDP the solution to the problem is the same bellman equation equation derived previously. We just substitute the new belief transition function and we get the corresponding belief bellman Equation, 2.15.

$$V^*(b_t) = r(b_t) + \gamma \max_{a_t} \int_{b_{t+1}} p(b_{t+1}|b_t, a_t)\, V^*(b_{t+1})\; db_{t+1} \tag{2.15}$$

Using this equation in this form is problematic, we are integrating over the space of beliefs and the transition function is a probability distribution over beliefs. The key to overcome this problem is to realise that if we know what the current measurement and applied action are there is only one valid possible belief. Thus the integration over beliefs vanishes. This can be seen by substituting the belief transition function, Equation 2.16, into the bellman equation Equation 2.15.

$$p(b_{t+1}|b_t, a_t) = \int_{y_t} p(b_{t+1}|b_t, a_t, y_t)\, p(y_t|y_{0:t-1}, a_{0:t})\; dy_t \tag{2.16}$$

After the substitution and re-arrangement of the sums we get an integral over all future value functions weighted by there probability, see Equation 2.17 which is the section of the bellman equation after the max. Since the observation is known (because the outer integral is over y_t), the integral over the beliefs vanishes since there is only one possible future belief which is given by the Bayesian filter function $\tau(b_t, a_t, y_t)$.

$$\gamma \max_{a_t} \int_{y_t} \underbrace{\left(\int_{b_{t+1}} p(b_{t+1}|b_t, a_t, y_t)\, V^*(b_{t+1})\; db_{t+1} \right)}_{1 \cdot V^*(\tau(b_t, a_t, y_t))} p(y_t|y_{0:t-1}, a_{0:t})\; dy_t \tag{2.17}$$

In the final belief state bellman equation, the integral of the belief state is replaced by an integration over observations, Equation 2.18.

$$\begin{aligned} V^*(b_t) &= r(b_t) + \gamma \max_{a_t} \int_{y_t} p(y_t|y_{0:t-1}, a_{0:t})\, V^*(f(b_t, a_t, y_t))\; dy_t \\ &= r(b_t) + \gamma \max_{a_t} \mathbb{E}_{y_t}\{V^*(f(b_t, a_t, y_t))\} \end{aligned} \tag{2.18}$$

The belief bellman equation is intuitive, the value of the current belief is the immediate reward plus the value of the future belief states weighted by the probability of a measurement which would result in these future belief states. It turns out that computing a value function using the above bellman function is not computationally tractable. Solving a POMDP problem as for the MDP case consists of finding the optimal value function from which the optimal policy can be derived. Essentially the same dynamic programming and reinforcement learning techniques can be applied to solve this problem. An exact solution is however only feasible when considering a finite state, action and observation space and a finite planning horizon T. Most early techniques for solving POMDPs used value iteration. It has been shown (Richard D. Smallwood (1973)) that because the reward function uses a linear operator (the expectation) and that the bellman backup operation (applying the bellman equation to the current value function) preservers the linearity, the value function after each updates is piece wise linear and continuous (PWLC). The intractability comes from the successive applications of the bellman backup operation which result in an exponential time and space complexity with respect to the planning horizon. A good text on the implementation of exact value iteration for POMDPs can be found (Thrun et al., 2005, Chap. 15) and here Kaelbling et al. (1998).

2.3 State of the art

Hansen (1998)

Tree search

Planning

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

Optimal control

$b = (\mu, \Sigma)$

?, **?**, Platt et al. (2010)

Optimal control methods represent the belief by a Gaussian function

Martinez-Cantin et al. (2009),**?**,Thrun et al. (2005)

Hauser (2010)

Ross et al. (2008)

He et al. (2011)

2.4 Summary

References

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

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. 1.1

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. ISBN 978-3-642-17451-3. 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

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artif. Intell.*, 101(1-2):99–134, May 1998. ISSN 0004-3702. doi: 10.1016/S0004-3702(98)00023-X. URL http://dx.doi.org/10.1016/S0004-3702(98)00023-X. 2.2.1

Ruben Martinez-Cantin, Nando de Freitas, Eric Brochu, JosÃľ 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

Edward J. Sondik Richard D. Smallwood. The optimal control of partially observable markov processes over a finite horizon. *Oper. Res.*, 21(5):1071–1088, October 1973. ISSN 0030-364X. doi: 10.1287/opre.21.5.1071. URL http://dx.doi.org/10.1287/opre.21.5.1071. 2.2.1

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

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005. ISBN 0262201623. 2.2.1, 2.3

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