PHD THESIS PROJECT

Probabilistic Behavioral Planning for Self-driving Vehicles

PhD Candidate: Ugo LECERF, INSA Lyon, Genie Electrique
Industrial partner: Renault Software Labs, Sophia Antipolis
Laboratory: EURECOM, Sophia-Antipolis

1 CONTEXT AND OBJECTIVES

Autonomous vehicles (AVs) could bring great benefits to society, from reducing road fatalities and injuries, to drastically reducing the carbon footprint of transportation systems, to providing independence to those unable to drive. Further, AVs offer the AI community many high-impact research problems in diverse fields including: computer vision, probabilistic modelling, pedestrian and vehicle modelling and multi-agent decision making, to name a few.

AV systems are typically built of a pipeline of individual components, linking sensor inputs to motor outputs. Raw sensory input is first processed by object detection and localization components, resulting in scene understanding. Scene understanding can then be used by a scene prediction component to anticipate other vehicles' motions. Finally, decision components transform scene predictions into commands that instruct AVs trajectories and short-term movements.

In the context of this Thesis project, the focus is on the development of new methodologies to enable decision components to operate an AV, based on its current understanding of the surrounding *environment*, while taking into account the probabilistic – and thus uncertain – nature of the problem, such that safety considerations and objectives can be systematically fulfilled. In other words, the focus of this Thesis is on the topic of **probabilistic reinforcement learning**: as a branch of machine learning, reinforcement learning (RL) is a computational approach to learning from interactions with the surrounding world and is concerned with

sequential decision making in unknown environments to achieve high-level goals. Usually, no sophisticated prior knowledge is available and all required information to achieve the goal has to be obtained through trials.

In a typical RL setup, an agent interacts with its surrounding world by taking *actions*. In turn, the agent perceives sensory input that reveal some information about the state of the world. Moreover, the agent perceives a *reward/penalty signal* that reveals information about the quality of the chosen action and the state of the world. The history of taken actions and perceived information gathered from interacting with the world forms the agent's *experience*. As opposed to supervised and unsupervised learning, the agent's experience is solely based on former interactions with the world and forms the basis for its next decisions. The agent's objective in RL is to find a sequence of actions, a *strategy*, that optimizes an expected long-term reward/cost. Solely describing the world is therefore insufficient to solve the RL problem: the agent must also decide how to use the knowledge about the world in order to make decisions and to choose actions. Since RL is inherently based on collected experience, it provides a general, intuitive, and theoretically powerful framework for **autonomous learning and sequential decision making under uncertainty**.

In the context of this Thesis proposal, the problem setting is inherently **hierarchical**. In a nutshell, the state of the world (also called the world model) is a probabilistic representation of the surroundings of an AV, including object dynamics; the agent must learn a trajectory, which is a composition of elementary actions. To make the problem more tractable, it is decomposed into a hierarchy of sub-tasks. At the high level, a *route planner* defines a high level goal, which is decomposed into medium and short term sub-goals. At the medium-term, a *behavioral planner* defines macro-actions, whereas at the low-level, a *motion planner* implements fine-grained actions involving, for example, acceleration/deceleration and steering angles. Hierarchical reinforcement learning brings several advantages: for example, policies learned to solve sub-problems can be reused for multiple parent tasks. In addition, the overall *value function* (which guides the agents behavior) can be represented more compactly, as the sum of separate terms that only depend on a subset of the state variables. This compact representation requires less data to learn, leading to more efficient and fast learning. The literature on hierarchical reinforcement learning also brings up several challenges (which we briefly review below): in particular, in this Thesis project, we will address the problem of learning how to decompose the overall goal into sub-problems, including the ability to factor in country-specific requirements and behaviors.

1.1 Objectives

We now clearly identify the objectives of this Thesis proposal:

1. The first objective we consider is how to **integrate the notion of uncertainty** in the behavioral and motion planning tasks. This involves the design of Bayesian, probabilistic reinforcement learning methods, with the underlying hypothesis that uncertainty permeates the full pipeline used to operate an AV, starting from the world model, the transition model up to the actions and rewards.

2. Given the hierarchical nature of the overall route planning problem, an open question

is how to **learn a good abstract representation** for actions at different levels of the hierarchy, instead of relying on hand-crafted rules as it is frequently done in the literature. Low-level actions are determined by the motion planner, whereas high-level goals are determined by the route planner. In the middle, the objective of the behavioral planner is to operate on a set of discrete actions that bridge the gap between high-level and low level goals.

3. The problem we consider in this Thesis project can be seen as a **multi-objective optimization** problem. Our objective, then, is to optimize for several aspects, including: i) **efficiency** (e.g., time to reach a position, fuel consumption, ...); ii) **safety** (e.g., taking into account regulatory guidelines); and iii) **comfort** (e.g., minimizing jerk, computed as the derivative of the acceleration). As a consequence, we will investigate the notion of a Pareto front of such an optimization problem, which allow exploring the tradeoff between such objectives.

2 Background

This Section is devoted to an overview of the literature, which we organize according to three main categories: probabilistic machine learning, Bayesian reinforcement learning, and Hierarchical reinforcement learning.

2.1 Probabilistic Machine Learning

Broadly speaking, in Bayesian learning, we make inference about a random variable X by producing a probability distribution for X. Inferences, such as point and interval estimates, may then be extracted from this distribution. Let us assume that the random variable X is hidden and we can only observe a related random variable Y . Our goal is to infer X from the samples of Y . A simple example is when X is a physical quantity and Y is its noisy measurement. Bayesian inference is usually carried out in the following way:

- We choose a probability density $P(X)$, called the *prior* distribution, that expresses our beliefs about the random variable X before we observe any data.
- We select a statistical model $P(Y|X)$ that reflects our belief about Y given X. This model represents the statistical dependence between X and Y.
- We observe data $Y = y$.
- We update our belief about X by calculating its posterior distribution using Bayes rule:

$$P(X|Y = y) = \frac{P(y|X)P(X)}{\int P(y|X')P(X')dX'}$$

Assume now that $P(X)$ is parameterized by an unknown vector of parameters θ in some parameter space Θ; we denote this as $P_\theta(X)$. Let $X_1, \cdots, X_n$ be a random i.i.d. sample drawn from $P_\theta(X)$. In general, updating the posterior $P_\theta(X|Y = y)$ is difficult, due to the need to

compute the normalizing constant at the denominator. However, for the case of conjugate family distributions, we can update the posterior in closed-form by simply updating the parameters of the distribution: this is a simple example of probabilistic nonparametric models such as Gaussian Processes (GPs), which we will extensively use in this Thesis proposal.

Probabilistic nonparametric models [9] offer flexibility and accurate quantification of uncertainty. These models are probabilistic in the sense that they interpret data as realizations of some unknown probabilistic process, and they use the language of probabilities (mainly Bayes theorem) to âĂIJreverse-engineerâĂİ it. The nonparametric aspect of these models is that they model the relationships among variables to describe the generative process by means of probability distributions over functions. Despite the infinite dimensional nature of these distributions, it is possible to tractably deal with the computations associated with the inference of these models. The combination of the probabilistic and nonparametric nature of these models is that the result of the inference process is a distribution over functions, which characterizes all functions that are compatible with the observed data. This is of fundamental importance to account for this source of uncertainty when making predictions and analyzing data [9].

While probabilistic nonparametric models offer powerful tools for learning from data, they are extremely complex to use in practice due to their poor scalability with the number of observations. This is due to the need to repeatedly solve hard algebraic problems involving large dense matrices. Some recent contributions demonstrate that it is possible to tackle these issues by combining ideas from statistical physics, probabilistic modeling, and algebra [4], but they also indicate that more work is needed to develop fully scalable solutions that do not introduce bias in predictions and quantification of uncertainty.

2.2 Bayesian Reinforcement Learning

Reinforcement learning (RL) [2, 10] is a class of learning problems in which an agent (or controller) interacts with a dynamic, stochastic, and incompletely known environment, with the goal of finding an action-selection strategy, or policy, to optimize some measure of its long-term performance. The interaction is conventionally modeled as a Markov Decision Process (MDP), or if the environment state is not always completely observable, as a partially observable MDP (POMDP).

In contrast to supervised learning methods that deal with independently and identically distributed (i.i.d.) samples from a given learning domain, the Reinforcement Learning approach involves agents learning from samples that are collected from the trajectories[1] generated by its sequential interaction with a given system (e.g., the environment generated by an egocentric view of an autonomous car).

Traditionally, RL algorithms have been categorized as being either *model-based* or *model-free*. In the former category, the agent uses the collected data to first build a model of the domain's dynamics and then uses this model to optimize its policy. In the latter case, the agent directly learns an optimal (or good) action-selection strategy from the collected data.

[1] The attentive reader shall not confuse the term "trajectory" used in this section as representative of the random path taken when exploring the underlying Markov Decision Process modeling the agent-system interaction, and the term "trajectory" referred to that taken by an autonomous vehicle on the road.

A major challenge in RL is in identifying good data collection strategies, that effectively balance between the need to explore the space of all possible policies, and the desire to focus data collection towards trajectories that yield better outcome (e.g., greater chance of reaching a goal, or minimizing a cost function). This is known as the *exploration-exploitation tradeoff* which, in other words, explains the tension that exists between either taking actions that are most rewarding according to the current state of knowledge, or taking exploratory actions, which may be less immediately rewarding, but may lead to better informed decisions in the future. This challenge arises in both model-based and model-free RL algorithms.

Bayesian reinforcement learning (BRL) is an approach to RL that leverages methods from Bayesian inference to incorporate uncertainty information into the learning process [8]. It assumes it is possible to express prior information about the problem in a probabilistic manner, and that new information can be incorporated using standard rules of Bayesian inference. BRL enjoys three main advantages, when compared to traditional methods:

1. A major advantage of the BRL approach is that it provides a principled way to tackle the exploration-exploitation problem. Indeed, the Bayesian posterior naturally captures the full state of knowledge, subject to the chosen parametric representation, and thus, the agent can select actions that maximize the expected gain with respect to this information state. Indeed, Bayesian methods applied to RL deal with this difficult problem by explicitly quantifying the value of exploration, which is made possible by maintaining a distribution over the "so-called" probability kernel, which models state transitions based on selected actions.

2. Another major advantage of BRL is that it implicitly facilitates regularization. As discussed in Section 2.1, for the general case, a Bayesian approach to inference induces an implicit regularization, which favors simple models and helps mitigating overfitting.

3. Finally, another advantage of adopting a Bayesian view in RL is the principled Bayesian approach for handling parameter uncertainty. Indeed, the goal is to explicitly represent uncertainty over the model parameters, such as the probability kernel and the reward function. One way to think about the Bayesian approach is to see the parameters as unobservable states of the system, and to cast the problem of planning in an MDP with unknown parameters as planning under uncertainty using the POMDP formulation.

Of course, several challenges arise in applying Bayesian methods to the RL paradigm. First, there is the challenge of selecting the correct representation for expressing prior information in any given domain. In the context of this Thesis project, in particular, the idea is also to encode information about safety rules: it is yet to be proven that such information can be succesfully incorporated in the prior, or if safety rules will override agents' decision in presence of large values of uncertainty. Second, defining the decision-making process over the information state is typically computationally more demanding than directly considering the natural state representation. It is then precisely in this context that we foresee important contributions of this Thesis project: namely, we will leverage approximation techniques, such as low-rank approximations, inducing points, and random feature projections, as well as a variational approach to Bayesian inference (as discussed in Section 2.1), to make BRL computationally efficient.

In conclusion, Bayesian reinforcement learning offers a coherent probabilistic model for reinforcement learning. It provides a principled framework to express the classic exploration-exploitation dilemma, by keeping an explicit representation of uncertainty, and selecting actions that are optimal with respect to a version of the problem that incorporates this uncertainty [7].

2.3 Hierarchical Reinforcement Learning

Despite the exponential growth of advances in reinforcement learning, one original shortcoming that has been extensively studied in the literature relates to the problem of a fully satisfactory method for incorporating hierarchies into reinforcement learning algorithms.

Many researchers (a noteworthy example is T. Dietterich [5]) have experimented with different methods for hierarchical reinforcement learning and hierarchical probabilistic planning. Previous research – extensively surveyed in works such as [1, 11, 10] – shows that there are several important design decisions that must be made when constructing a hierarchical reinforcement learning system. The problems we shortly overview below still constitute open issues as of today, and will be addressed in the development of this Thesis proposal.

The first issue is *how to specify sub-tasks*. Hierarchical reinforcement learning involves breaking the target Markov decision problem into a hierarchy of sub-problems. There are three general approaches to defining these sub-tasks. One approach is to define each sub-task in terms of a fixed policy that is provided by the programmer, based on domain knowledge. The second approach is to define each sub-task in terms of a non deterministic finite-state controller. This allows the programmer to design a "partial-policy" that constraints the set of actions available, but does not specify a complete policy for each sub-task. A third method consists in defining each sub-task in terms of a termination predicate and a local reward [5].

The second design issue is whether to employ state abstractions within sub-tasks. A sub-task employs state abstraction if it ignores some aspects of the state of the environment. An example of approach that explicitly addresses this issue is the MAX-Q method [5].

The third design issue concerns the non-hierarchical "execution" of a learned hierarchical policy. Ordinarily, in hierarchical reinforcement learning, the only states where learning is required at the higher levels of the hierarchy are states where one or more of the subroutines could terminate (plus all possible initial states). But to support non-hierarchical execution, learning is required in all states (and at all levels of the hierarchy). In general, this requires additional exploration as well as additional computation and memory.

The fourth issue is what form of learning algorithm to employ. An important advantage of reinforcement learning algorithms is that they typically operate online. However, finding online algorithms that work for general hierarchical reinforcement learning has been difficult, particularly within the termination predicate family of methods. Due to the context of this Thesis project proposal, online learning is not necessarily the best method to use. For this reason, this fourth issue is not going to be problematic, although several research approaches successfully address it [5].

3 METHODOLOGY

In this section we outline the methodology that will be adopted throughout the execution of this PhD project proposal:

- **Bibliographic study**: The proposed research topic contains a real academic dimension, and a thorough bibliographic study is crucial to address it. First, the PhD student will have to familiarize with the concept of Bayesian inference [3] and probabilistic machine learning, with the goal of being able to come up with methodological contributions in this field. More precisely, the PhD student will investigate applications of probabilistic inference in the reinforcement learning domain, for which the literature has been expanding in the last years [8, 7]. Given the application context of this Thesis proposal, it is important to study the state of the art in motion planning algorithms as well, and how to integrate the new methodologies to arrive at the definition of probabilistic trajectories. In addition, a patent anticipation search will be conducted. All of these elements will guide the further work.

- **Methodological contributions**: the goal of Thesis proposal is to go beyond the state of the art in motion planning, by incorporating the notion of uncertainty that permeates both the data and the learned statistical models. Due to the extremely demanding computational requirements of Bayesian inference [3, 9, 4], we expect to come up with methodological contributions in approximation algorithms to render probabilistic behavioral planning feasible in practice [7]. In addition, we will focus on Hierarchical reinforcement learning, starting from a solid literature exploration []. Finally, the industrial environment of this Thesis proposal, which include the need to develop solutions for consumer vehicles (as opposed to high-end vehicles) as well as the requirement for developing methodologies that can seamlessly integrate with regulatory-issued safety constraints, calls for methodological contributions that comply with such operational constraints.

- **Simulation environments**: the techniques suggested in this Thesis Proposal, namely those that belong to the large family of reinforcement learning, require an appropriate simulation environment to be effectively exploited. In fact, reinforcement learning alone cannot be directly used by an autonomous vehicle to be trained âĂIJin the wildâĂİ, because the training process follows a simple âĂIJtrial-and-errorâĂİ approach that would result in catastrophic losses of vehicles. As a consequence, a first step to validate the methodological contributions outlined above consists in using realistic simulation environments, whereby probabilistic motion planning based on reinforcement learning will operate. This step will be approached both using virtual environments (a car simulator, such as CARLA [6] for example) and real-life traces provided by Renault Software Labs. Such traces should be representative of, for example, ground markings, traffic signs, navigable space, obstacles, moving objects and their attributes (position, longitudinal, lateral speeds, confidence, object class). A second step to validate the methods proposed in this Thesis project is to enhance the current reinforcement learning paradigm by explicit human feedback. This would allow, when the project reaches acceptable maturity

levels, to bypass the lengthy and computationally demanding simulation process, and pave the road for more realistic experiments which are described next.

- **Real-life experiments**: the ultimate goal of this project is to contribute to both the academic literature and the industrial sector with a through experimental evaluation of probabilistic motion planning methods. Building on the possibility to include a human-in-the-loop in the reinforcement learning methodology, we will develop methods to perform real-life test-fields using the expertise, the infrastructure and the available vehicles from Renault Software Labs.

4 Research Environment

The supervision of the PhD student will be arranged according to both academic and industrial objectives of the Thesis:

- **EURECOM**: this is an internationally renowned team founded by Prof. Pietro Michiardi and Prof. Maurizio Filippone, who are experts in probabilistic machine learning, and in the design and implementation of scalable methodologies for large-scale learning problems. In the context of this Thesis, EURECOM will provide the necessary background and mathematical tools to contribute with novel methodologies in the context of probabilistic reinforcement learning. More precisely, the PhD student will be exposed to Bayesian inference techniques for a sound quantification of uncertainty that permeates the learning process, and the required approximation methodologies to approach probabilistic reinforcement learning with computationally efficient algorithmic implementations.

- **Renault Software Labs**: this is a department of the Renault Alliance dedicated to the design, prototyping and productization of autonomous driving solutions in which Dr-Eng. Sébastien Aubert and Eng. Philippe Weingertner are contributing. In the context of this Thesis, RSL will provide theoretical knowledge related to partially observable Markov decision processes and reinforcement learning. In order to fit industrial needs, the PhD student will challenge the superiority of its contributions with synthetic data - based on in-house end-to-end simulator - and realistic data - remote and on-board -.

The student will benefit from an active and stimulating research environment, in an international scientific context, and will be brought to present his work in the major conferences of the research fields related to this project (Bayesian statistics, machine learning theory, reinforcement learning, etc.). The expected thesis work may lead to scientific publications (newspapers, conferences), potentially accompanied by patents and technical reports. These phases will ensure steady progress of the PhD student work, as well as prepare the road for writing the final PhD Thesis manuscript.

REFERENCES

[1] Andrew G. Barto and Sridhar Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13(1-2):41–77, January 2003.

[2] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

[4] Kurt Cutajar, Edwin V. Bonilla, Pietro Michiardi, and Maurizio Filippone. Random feature expansions for deep gaussian processes. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 884–893, 2017.

[5] Thomas G. Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *J. Artif. Int. Res.*, 13(1):227–303, November 2000.

[6] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[7] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *Found. Trends Mach. Learn.*, 8(5-6):359–483, November 2015.

[8] Mykel J. Kochenderfer, Christopher Amato, Girish Chowdhary, Jonathan P. How, Hayley J. Davison Reynolds, Jason R. Thornton, Pedro A. Torres-Carrasquillo, N. Kemal Üre, and John Vian. *Decision Making Under Uncertainty: Theory and Application*. The MIT Press, 1st edition, 2015.

[9] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[10] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[11] Marco Wiering and Martijn van Otterlo. *Reinforcement Learning: State-of-the-Art*. Springer Publishing Company, Incorporated, 2014.