THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Risk-Averse Decision-Making under Parametric Uncertainty

Hannes Eriksson

Department of Computer Science and Engineering
Chalmers University of Technology
Gothenburg, Sweden, 2022

Risk-Averse Decision-Making under Parametric Uncertainty

Hannes Eriksson

ISSN 1652-876X
Technical Report No.
Department of Computer Science and Engineering
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Phone: +46 (0)31 772 1000
www.chalmers.se

Printed by Chalmers Reproservice
Gothenburg, Sweden, May 2022

Abstract

For sequential decision-making problems with potentially catastrophic consequences appropriate risk assessment may be required. In contrast to traditional techniques for decision-making under uncertainty that aim to maximise performance in expectation, we chose to focus on other properties of the probability distribution. For instance, in an application such as autonomous driving, the chance of causing an accident might be small but yet fatal. A decision-maker focusing on performance in the worst outcomes may be able to obtain a safer decision-making process by keeping this in mind. We propose frameworks for quantifying uncertainty under the reinforcement learning framework and develop algorithms that allow for risk-sensitive decision-making under uncertainty.

Keywords: Reinforcement learning, autonomous driving, risk-sensitive learning, uncertainty estimation.

List of Publications

This thesis is based on the following publications:

[A] **Hannes Eriksson**, Christos Dimitrakakis, "Epistemic Risk-Sensitive Reinforcement Learning". Published in European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2020.

[B] Emilio Jorge, **Hannes Eriksson**, Christos Dimitrakakis, Debabrota Basu, Divya Grover, "Inferential Induction: A Novel Framework for Bayesian Reinforcement Learning". Published in Proceedings of Machine Learning Research Volume 137 (PMLR), 2021.

[C] **Hannes Eriksson**, Debabrota Basu, Mina Alibeigi, Christos Dimitrakakis, "SENTINEL: Taming Uncertainty with Ensemble-based Distributional Reinforcement Learning". Accepted for Association for Uncertainty in Artificial Intelligence (UAI), 2022.

Other publications by the author, not included in this thesis, are:

[D] **Hannes Eriksson**, Christos Dimitrakakis, Lars Carlsson, "High-Dimensional Near-Optimal Experiment Design for Drug Discovery via Bayesian Sparse Sampling". *arXiv preprint arXiv:2104.11834, 2021.*

[E] **Hannes Eriksson**, Debabrota Basu, Mina Alibeigi, Christos Dimitrakakis, "Risk-Sensitive Bayesian Games for Multi-Agent Reinforcement Learning under Policy Uncertainty". *Included in The 13th Workshop on Optimization and Learning in Multiagent Systems @ AAMAS'22.*

Acknowledgments

I want to thank all the people in the Wallenberg AI, Autonomous Systems and Software Program (WASP) for all the interesting times we have had during travels, courses and conferences. Likewise, I would like to thank the people I have interacted with at Chalmers University of Technology during my stay here. I would also like to thank my colleagues at Zenseact AB for all the interesting times and challenges related to our work in autonomous driving. I would like to give special thanks to the people at Harvard SEAS who were involved with my stay there before the inception of my PhD studies, these were, David, Paul, Goran, Rafael and more, who inspired me to work on my PhD. I would also like to give extra thanks to the people involved with the advanced graduate program at Zenseact AB, for fostering a great community for research related to autonomous driving, these are Mats, Carl and more. I would also like to thank my supervisors, these are, Alexander, Nasser, who helped guiding me in the early parts of this PhD work, Mina, who has been guiding me for the latter parts of the projects and has shown great commitment and interest in our work. I would like to give special thanks to Devdatt, who initially suggested for me to apply for a PhD position, for being examiner both for my master thesis and PhD thesis and for his role in the Data Science division at Chalmers University of Technology. I would also like to thank all my current and past collaborators, these are, Christos, Debabrota, Divya, Emilio, Mina and Tommy. I would like to give thanks to the members of the research group under Christos, which now span multiple countries, these are (former included) Aristide, Divya, Emilio, Debabrota, who I have had the pleasure of working with and discussing over all these years, as well as Thomas, Meirav and Ann-Marie in our weekly discussions. Finally, I would like to thank Christos who has endured all these years with me, from my master thesis work, to my work as project assistant, to my research trip to Boston and throughout the PhD studies. Without his assistance and guidance this would never have been possible.

Contents

Part I

Overview

CHAPTER 1

Introduction

Sequential decision-making processes involving uncertainties about transition process and the objective may require some careful examination of the particular involved uncertainties. One application where this is of particular interest is in the *Autonomous Driving* (AD) setting, which will be discussed further in this chapter. Following that, the framework under which this decision-making process is studied is introduced, which is *Reinforcement Learning* (RL). Finally, we discuss the contributions of our work and the outline of this thesis.

1.1 Autonomous Driving

Part of the AD pipeline includes a sequential decision-making process whereby an autonomous agent is tasked with selecting appropriate *actions* to fulfil some pre-specified goal. These actions could relate directly to physical inputs to the vehicle, such as controlling the throttle or the steering of the vehicle, or they could involve more abstract decisions such as `follow the vehicle ahead`, `overtake the vehicle ahead`, etc. The notion of a goal, or *objective*, is directly related to the task at hand, where for instance, the goal of an agent tasked with driving through an intersection has the specified goal of ending

up on the other side of the intersection. During the task, the agent might also have a set of sub-goals, which could involve adhering to the traffic rules, driving at a comfortable pace and being mindful of other traffic participants.

1.2 Uncertainty in Autonomous Driving

Typically, the agent does not have complete information of the task at hand, notably the agent might only have access to some high level objective such as *arrive at some pre-specified location as quickly as possible, while adhering to the traffic rules* while outfitted with sensors such as cameras, radar, GPS, map information, etc. The agent then, using the information available from these sensors, has to interpret the world and estimate the current *state of the world* from which it is supposed to act from. Part of the problem revolves around the uncertainty about the evolution of the decision-making process where the agent has to guess what will happen in the future in order to make the correct decisions in the present.

1.3 Reinforcement Learning

One framework for handling *decision-making under uncertainty* is RL, which has seen great success [1]–[3], and is something that has been studied extensively for the field of AD as well [4], [5]. Part of the RL framework involves the construction of a *Markov Decision Process* (MDP) [6] which describes how the state of the world evolves, which actions the agent can take and the goodness, or *reward*, associated with it. Typically, the true underlying MDP is unknown and the agent has to estimate this MDP from available data. This introduces a sort of uncertainty related to the knowledge available to the agent, henceforth to be called *epistemic uncertainty* [7]–[10] . This exists in contrast to another kind of uncertainty, which is inherent to the MDP is termed *aleatory uncertainty* [11]. Aleatory uncertainty is abundant in applications with high stochasticity, such as games of chance. In applications such as for autonomous driving, with mostly deterministic mechanics, this source of uncertainty might not be so great, given that world dynamics are known. These two kinds of uncertainties form the basis of this thesis and the differences, applications and importance of them will be stressed throughout this thesis.

1.4 Contributions

In Eriksson and Dimitrakakis [9] we develop and introduce a risk-sensitive Bayesian RL framework for decision-making under *epistemic* uncertainty for discrete and continuous state space RL problems. In addition to that, we propose two algorithms, one based on approximate dynamic programming and one based on the Bayesian policy gradient.

In the work Jorge *et al.* [12] we introduce a novel framework for Bayesian distributional RL by appropriately marginalising out the variables in such a way that three new approaches can be formulated. We propose one of them, Bayesian Backwards Induction and demonstrate its performance in the paper.

Lastly, in Eriksson *et al.* [10] we propose a novel risk measure, termed *composite risk*, which takes into account both aleatory and epistemic uncertainty and appropriately weights them together. We prove superiority over previous methods of joining the risk measures theoretically and propose an ensemble-based algorithm that can quantify this new risk measure.

1.5 Thesis Outline

The thesis is initiated with a chapter covering the main ingredients the included publications are based upon, in Chapter 2. These include the basics of RL and the constructions which allow for risk-averse decision-making in RL. In the next chapter, Chapter 3, the value function representations studied in papers [B, C] are elaborated upon. In Chapter 4, the papers [A, C] are discussed, and the topic introduced in Chapter 3 is used to aide risk-sensitive decision-making by constructing robust value function distributions. Finally, in Chapter 5 the thesis is wrapped up and future works are discussed.

CHAPTER 2

Background

In this chapter we will cover the necessary background information. Much of the work concern estimating or constructing a belief over MDPs. In Section 2.1 we cover the basics of MDPs and the optimisation over known MDPs. In the following Section 2.2, we elaborate on the case when the MDP is not known.

2.1 Dynamic Programming

In this section, we go over the fundamentals of using Dynamic Programming (DP) to solve MDPs.

Definition 1 (Markov Decision Process)**:** *A Markov Decision Process μ is a tuple $\mu = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$, where $\mathcal{S} = \mathbb{R}^d$ is a d-dimensional representation of the state, $\mathcal{A}$, the permissible action set of size M, $\mathcal{R} = \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function associating the goodness of taking an action a for a particular state s, $\mathcal{T} = \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ a transition kernel, describing the evolution of the Markov process depending on the current state and selected action and finally γ, a discount factor which determines the effective horizon of the problem.*

In addition to the formalism surrounding the MDP, we introduce a couple of important concepts such as the *policy* $\pi \in \Pi$, which describes the strat-

egy of the agent. The policy for a particular state s describes a probability distribution over actions $\pi = \mathbb{P}(a_t \,|\, s_t)$. Furthermore, the transition kernel $\mathcal{T} = \mathbb{P}_\mu(s_{t+1} \,|\, s_t, a_t)$ as previously mentioned, describes how the stochastic process evolves over time. The reward function $\mathcal{R} = \mathbb{P}_\mu(r_{t+1} \,|\, s_t, a_t)$ describes instead the goodness of selecting action a in state s. These concepts together allow us to define another important concept in RL, namely *value functions*.

Definition 2 (Value function)**:** *The value function $V_\mu^\pi(s)$ describes the expected utility of being in state s, for MDP μ, following policy π.*

$$V_\mu^\pi(s) = \mathbb{E}_\mu^\pi\Big[\sum_{t=0}^{\infty} \gamma^t r_t \,|\, s_0 = s\Big], \tag{2.1}$$

where $\gamma \in (0, 1]$ is a discount factor, determining the effective horizon of the problem.

This is the most common definition of value function, that is the expected sum of discounted future rewards. Other alternative definitions could involve the infinite horizon average reward or one that takes risk into account.

One goal in RL is to find the *optimal* value function $V_\mu^* \triangleq \max_\pi V_\mu^\pi$, and the associated *optimal* policy $\pi^* \triangleq \arg\max_\pi V_\mu^\pi$.

We can now define a function operator, termed the *Bellman operator*, $\mathscr{P}^\pi : V \to V$ as,

$$\mathscr{P}^\pi V(s) \triangleq \mathbb{E}_\mu^\pi[\mathcal{R}(s, a)] + \mathbb{E}_\mu^\pi\Big[\mathcal{T}(s, a)V(s')\Big]. \tag{2.2}$$

Iteratively applying $\mathscr{P}$ for all states $s \in \mathcal{S}$ for a particular MDP μ can be used to obtain the value function associated with the policy π, MDP μ and state s. Another operator of interest is the *Bellman optimality operator*, $\mathscr{P} : V \to V$, which can be defined as,

$$\mathscr{P}V(s) \triangleq \max_{a \in \mathcal{A}} \mathbb{E}_\mu^\pi[\mathcal{R}(s, a)] + \mathbb{E}_\mu^\pi\Big[\mathcal{T}(s, a)V(s')\Big]. \tag{2.3}$$

These operators has been shown in Bertsekas and Tsitsiklis [13] to be contraction mappings and thus, repeated applications of them will result in convergence to its corresponding value function, i.e., $\lim_{t\to\infty} \underbrace{\mathscr{P}^\pi(\ldots(\mathscr{P}^\pi}_{t \text{ times}} V)) = V_\mu^\pi$ and $\lim_{t\to\infty} \underbrace{\mathscr{P}(\ldots(\mathscr{P}}_{t \text{ times}} V)) = V_\mu^*$.

2.2 Reinforcement Learning

In the previous section we have assumed the reward function $\mathcal{R}$ and the transition kernel $\mathcal{T}$ to be known. In the case where they are known, we have efficient ways to compute the optimal value function. However, for the vast majority of interesting decision-making problems, the MDP is not known. Much of the theory of RL revolves around estimating a MDP from existing observations and using DP to compute the optimal value function given the current data. This introduces an important problem in RL known as the *exploration-exploitation dilemma*, where an agent must choose to either gather more information about the true underlying MDP by exploring, or acting optimally using its current knowledge. This dilemma will be further discussed later on in Chapter 3.

Bayesian Reinforcement Learning

One framework for engaging with decision-making problems with unknown MDP is through the Bayesian Reinforcement Learning (BRL) framework. Under this framework, we study distributions over MDPs in a probability space $(\mathcal{M}, \mathcal{F}, \xi)$, where $\mu \in \mathcal{M}$ is the set of admissible MDPs, $\mathcal{F}$ an appropriate $\sigma-$algebra and ξ a probability function over subsets of $\mathcal{M}$. The works [9], [12] rely on this framework. In [9] a Dirichlet product-prior is admitted over transition functions $\mathbb{P}(s_{t+1} \mid s_t, a_t)$ in the discrete case and independent Gaussian processes for the continuous case. The reward functions $\mathbb{P}(r_t \mid s_t, a_t)$ are Normal-Gamma and Gaussian process distributed for the discrete and continuous case, respectively. After constructing the appropriate beliefs over reward function $\mathcal{R}$ and transition kernel $\mathcal{T}$, a MDP $\mu \sim \xi$ can be sampled, and the approach described in Section 2.1 can be used to arrive at an optimal policy π^* given μ. The uncertainty about μ given by $\xi(\mu)$ is crucial, and its purpose is detailed further in Section 2.3.

Distributional Reinforcement Learning

Under the Distributional Reinforcement Learning (DRL) framework introduced by Bellemare *et al.* [14], with some similarities to earlier works on Bayesian Reinforcement Learning, such as [15], [16], involves representing either the return or value function as a random variable. This construction

allows for learning a robust representation of the return or value function distribution. At the time of publication, DRL agents such as Bellemare *et al.* [14] and later works such as Hessel *et al.* [17] displayed state-of-the-art performance on Deep RL tasks. In addition to the general performance benefit shown, this framework allows for a novel way of representing the uncertainty about the return or value function. This is crucial for research fields such as Risk-Sensitive Reinforcement Learning (RSRL), which is further elaborated on in Section 2.3.

2.3 Risk-Sensitive Reinforcement Learning

For general reinforcement learning problems, the objective is commonly maximisation of the *expected* return. For various applications, such as autonomous driving, where real-life accidents can have catastrophic consequences, it might be of value to consider other properties of the return distribution. For instance, in Eriksson and Dimitrakakis [9] an exponential utility function is used. This allows for a change of objective from the traditional risk-neutral one to the following,

$$U_\beta(\xi, \pi) \triangleq \frac{1}{\beta} \log \int_{\mathcal{M}} \exp\left(\beta \mathbb{E}^\pi_\mu[R]\right) \mathrm{d}\xi(\mu). \tag{2.4}$$

This formulation is based upon the objective stated by Mihatsch and Neuneier [11], which has some interesting properties. In general, the choice of $\beta < 0$ admits risk-averse decision-making and risk-seeking for $\beta > 0$. Another approach is studied in [10], where the maximisation is over a *conditional value-at-risk* (CVaR) objective, based upon work by Rockafellar, Uryasev, *et al.* [18]. In this case, the risk-sensitive objective for the aleatory case could be in the form,

$$CVaR_\alpha[R \,|\, \mu, \pi] \triangleq \mathbb{E}^\pi_\mu\Big[R \,|\, R \leq \nu_\alpha \wedge \mathbb{P}(R \geq \nu_\alpha) = 1 - \alpha\Big]. \tag{2.5}$$

The usage of CVaR adds the possibility of optimising the objective in the left-most or right-most tail of the distribution. The choice of $\alpha \in (0, 1]$ controls for up to which quantile the distribution should be considered.

A general discussion about the pertinence of utility functions and risk measures is discussed in Chapter 3.

CHAPTER 3

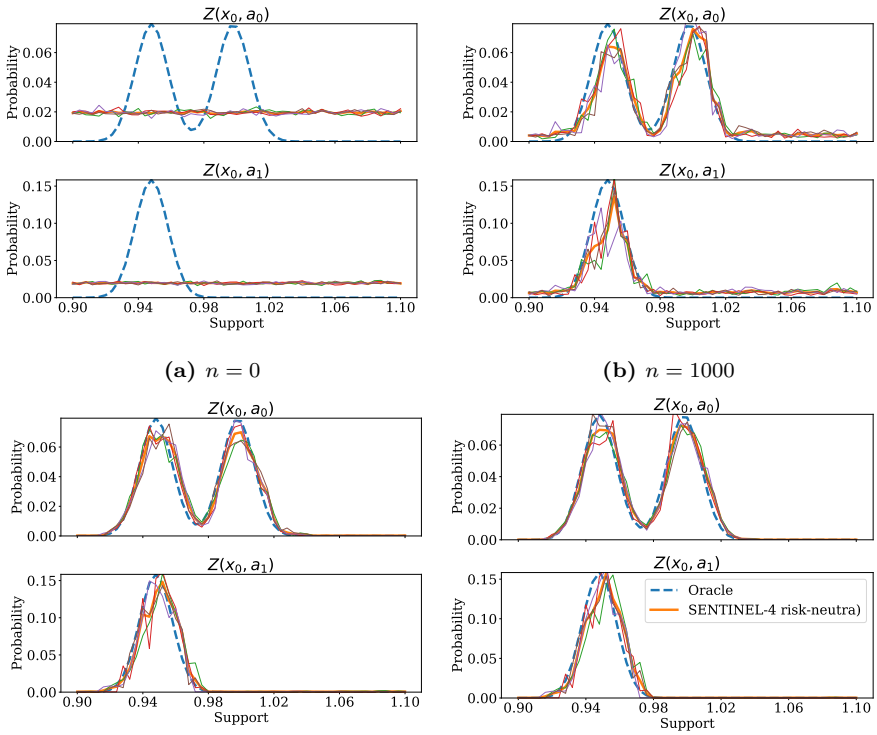
Value Function Representations under Uncertainty

Traditional RL techniques rely on accurate estimates of the value-function $V_{\mu}^{\pi}(s)$ in order to make decisions. For uncertainty aware applications such as risk-sensitive decision-making, where other properties of the return distribution is the main focus, it is important to have an accurate representation of the full distribution.

3.1 Value Function Representations using Bootstrapping

In the work Eriksson *et al.* [10] we adopt an approach whereby the uncertainty representations is made over an ensemble of CDQN agents [14]. A visualisation of using the proposed algorithm SENTINEL-K to learn a small toy problem can be seen in Figure 3.1. The dashed line indicates the actual return observed taking the chosen action a in the state x_0. The thick line is the estimated return distribution, marginalised over the 4 individual CDQN estimators. As can be seen, as more data is obtained ($n \gg 0$) the estimated return converges to the oracle return. The thin lines are each of the individual

estimators.



(a) $n = 0$ **(b)** $n = 1000$

(c) $n = 5000$ **(d)** $n = 10000$

Figure 3.1: Return distributions of a_0 and a_1 for $0, 1000, 5000$ and 10000 data points (n) respectively. The blue dashed line is the categorical approximation of $Z(s_0, a_0)$ and $Z(s_0, a_1)$ respectively. The thick orange line is the marginal posterior $\mathbb{P}(\hat{Z})$ with SENTINEL-4. The thin lines are the posteriors of the individual estimators.

Interestingly, as can be observed in both Figure 3.1 and earlier works such as Osband *et al.* [19], we can see that creating an ensemble of estimators can boost performance, as also indicated in [10]. The addition of multiple estimators not only allows us to consider the aleatory uncertainty (as represented by the distribution `for` each estimator), but we can also observe epistemic uncertainty by looking at the distribution `over` estimators. This is similar

to the works of [7], [8], [20] who all consider both the intra- and inter-model uncertainty, in various ways.

3.2 Bayesian Value Function Distributions

The main motivation in Jorge *et al.* [12] is the ability to represent a value function distribution $\mathbb{P}^{\pi}_{\beta}(V_t, \ldots, V_T \mid D)$ from time step t to T given some data D, a policy π and a belief β. By starting from T and going backwards to t, we can inductively compute the value function distribution $\mathbb{P}^{\pi}_{\beta}(V_i \mid D)$ in the following way,

$$\mathbb{P}^{\pi}_{\beta}(V_i \mid D) = \int_{\mathcal{V}} \mathbb{P}^{\pi}_{\beta}(V_i \mid V_{i+1}, D)\, \mathrm{d}\mathbb{P}^{\pi}_{\beta}(V_{i+1} \mid D). \tag{3.1}$$

The crucial difference between this approach and earlier approaches is that the distribution over MDPs $\mathbb{P}^{\pi}_{\beta}(\mu \mid V_{i+1}, D)$ includes information about the value function V_{i+1}. Using this framework, we devise a Monte Carlo method for estimating 3.1. This yields us a way of evaluating policies under this framework. The next step is to identify an approximately optimal policy, which is done using the algorithm *Bayesian Backwards Induction*. The full algorithmic details and results are available in Jorge *et al.* [12].

CHAPTER 4

Risk-Sensitive Reinforcement Learning under Parametric Uncertainty

In this chapter, we will discuss two approaches to doing risk-sensitive reinforcement learning. In the first approach, seen in Section 4.1, based upon Eriksson and Dimitrakakis [9], an exponential utility function is used to obtain risk-adjusted utilities. In the section, the two algorithms presented in the work are introduced. In the latter section given in Section 4.2, an approach based on coherent risk measures is discussed. In that section, we will go through the differences between that approach and the earlier approach, demonstrate why coherent risk measures are of interest and finally, propose an algorithm that effectively uses the proposed composite risk measure for decision-making under uncertainty.

4.1 Risk-Sensitive Reinforcement Learning with Exponential Utilities

In the work Eriksson and Dimitrakakis [9] an exponential utility function is used to allow for risk-adjusted utilities. In particular, the chosen form is based

upon the work by Mihatsch and Neuneier [11] which has the property that for maximisation of it leads to the following objective,

$$\frac{1}{\beta}\log \mathbb{E}[\exp(\beta Z)] = \mathbb{E}[Z] + \frac{\beta}{2}\mathrm{Var}[Z] + \mathcal{O}(\beta^2), \tag{4.1}$$

where it is clear that varying β leads to weighting the higher moments differently. For $\beta \to 0$, the risk-neutral objective is obtained, while for $\beta < 0$, higher variance and other moments are penalised. For risk-seeking behavior at $\beta > 0$ instead, then, e.g. higher variability will be premiered. An illustration of this phenomenon can be seen in Figure 4.1 where data sampled from 5 different normal distributions with the same mean is transformed using the proposed utility function. We can see that when, $U_{\beta\to 0} = 0$, which would correspond to risk-neutral behavior. Further, a risk-averse agent ($\beta < 0$), is maximised for the normal distribution with the lowest standard deviation, assuming all the means are equal. In contrast to this, the risk-seeking agent strictly prefers the normal distributions with higher standard deviation if the mean is kept intact.

Optimisation of Agents with Risk-Sensitive Utilities

In the work Eriksson and Dimitrakakis [9] as mentioned in Section 2.2, we use Dirichlet product-priors for the transition probabilities, and NormalGamma product-priors for the reward function,

$$\xi(\mu) \triangleq \mathbb{P}(\mu) = \prod_{s,a} \mathrm{Dir}(\theta_s^a) \times \mathrm{NG}(\vartheta_s^a) \tag{4.2}$$

More details on the parameters and the posterior update of the Normal-Gamma prior are given by Murphy [21]. From Eq. 4.2 it is clear that the addition of a belief over MDPs $\mu \in \mathcal{M}$ allows for the consideration of uncertainty about the MDP parameters. As explained in Section 2.1, if the MDP μ and policy π is fixed, we can obtain the value-function $V_\mu^\pi(s)$ for each state s. Since there is one corresponding value-function for each MDP μ if the policy is kept fixed, then if the belief over MDPs $\xi(\mu)$ is considered then naturally we obtain the value-function distribution $\mathcal{V}_\xi^\pi$ given by,

$$\mathcal{V}_\xi^\pi \triangleq \mathbb{P}(V_\mu^\pi \mid \mu \sim \xi, \pi, s_0 = s) \tag{4.3}$$

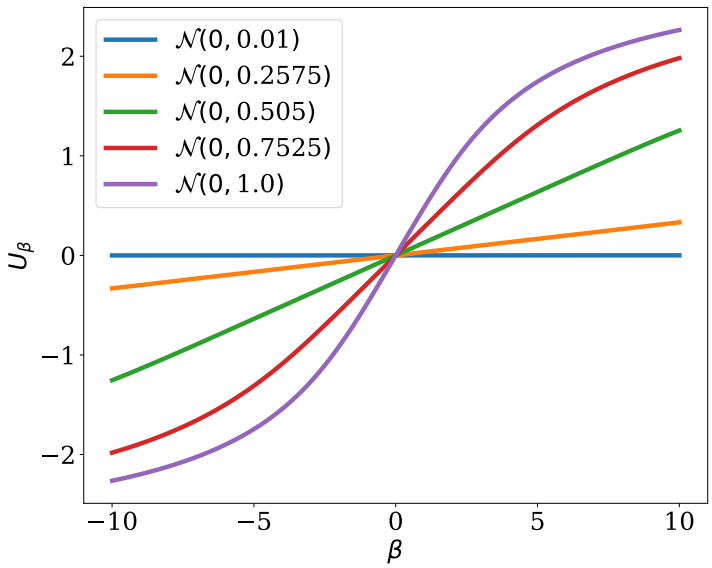


Figure 4.1: A small experiment demonstrating how the utility function proposed in Eq. 4.1 is impacted by changes in the risk-sensitive parameter β and by modifying the data distribution.

Algorithm 1 Epistemic Risk Sensitive Backwards Induction (ERSBI)

Input: $\mathcal{M}$ (set of MDPs), ξ (current posterior)
repeat
  for $\mu \in \mathcal{M}$ $s \in \mathcal{S}$, $a \in \mathcal{A}$ **do**
    $Q_\mu(s,a) = \mathcal{R}_\mu(s,a) + \gamma \sum_{s'} \mathcal{T}_\mu^{ss'} V_\mu(s')$
  end for
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    $\mathcal{Q}_\xi(s,a) = \sum_\mu \xi(\mu) U[(Q_\mu(s,a)]$
  end for
  $\pi(s) = \arg\max_a \mathcal{Q}_\xi(s,a)$.
  for $\mu \in \mathcal{M}$ **do**
    $V_\mu(s) = Q_\mu(s, \pi(s))$.
  end for
until *convergence*
return π

In the following sections, the uncertainty about the expected return $\mathbb{E}_\mu^\pi[R], V_\mu^\pi$ as $\mu \sim \xi$ and $\mathcal{V}_\xi^\pi$ are all considering the same thing, namely the uncertainty about the value function due to a probability distribution over MDPs.

A risk-averse decision-maker might seek to obtain a behavioural policy π^E that maximises performance for the risk-adjusted utilities as per Eq. 4.1. Such an agent would prefer policies that result in lower variability of values (keeping the expectation the same) to a risk-neutral agent. In Eriksson and Dimitrakakis [9] we first propose an algorithm for computing risk-sensitive policies using approximate dynamic programming, following the work of Dimitrakakis [22]. The proposed algorithm, ERSBI is an epistemically risk-sensitive backward induction approach using multiple models. The algorithm can be seen in Algorithm 1

Note that the proposed approach in Algorithm 1 is designed for discrete domains. For continuous domains we propose another algorithm based upon policy gradients, see work by Sutton *et al.* [23], where instead optimisation is done over the parameters of a *stochastic* policy. Deriving the policy gradient

Algorithm 2 Epistemic Risk Sensitive Policy Gradient (ERSPG)

Input: Policy parametrisation θ_t, β_t (current posterior).
repeat
  `Simulate to get` θ_{t+1}
  for $i = 1$ **to** N **do**
    $\mu_{(1)}, \mu_{(2)} \sim (\mathcal{M}_t, \mathcal{R}_t)$
    for $j = 1$ **to** M **do**
      $\tau^{(1)}_{\mu_{(1)}}, \tau^{(2)}_{\mu_{(1)}} \sim \pi_\theta, \mu_{(1)}$
      $\tau^{(3)}_{\mu_{(2)}} \sim \pi_\theta, \mu_{(2)}$
    end for
  end for
  $\theta_{t+1} \leftarrow \theta_t - [\sum_{i=0}^{N} \exp\left(\beta {\tau_{\mu_i}}^{(1)}\right) {\tau_{\mu_i}}^{(2)} \nabla_\theta \log \pi_\theta(a|s)] / [\sum_{i=0}^{N} \exp\left(\beta {\tau_{\mu_i}}^{(3)}\right)]$
  Deploy $\pi_{\theta_{t+1}}$ and obtain $\tau \sim \mu, \pi_{\theta_{t+1}}$
  $\xi_{t+1} \leftarrow \xi_t, \tau$
until *convergence*

update for the utility function in Eq. 4.1 leads to,

$$\nabla_\theta \frac{1}{\beta} \log \int_{\mathcal{M}} \exp\left(\beta \mathbb{E}_\mu^{\pi_\theta}[R]\right) d\xi(\mu) = \frac{\int_{\mathcal{M}} \exp\left(\beta \mathbb{E}_\mu^{\pi_\theta}[R]\right) \nabla_\theta \mathbb{E}_\mu^{\pi_\theta}[R] d\xi(\mu)}{\int_{\mathcal{M}} \exp\left(\beta \mathbb{E}_\mu^{\pi_\theta}[R]\right) d\xi(\mu)}. \tag{4.4}$$

Using the derived gradient update in Eq. 4.4 we propose an epistemically risk-sensitive policy gradient algorithm (ERSPG) which can be seen in Algorithm 2. The procedure uses utility rollouts using the sampled MDPs and reweights them appropriately using the exponential utility function before updating the parameters of the policy.

Experimental Results using Exponential Utilities

We leave the experimental results section to the paper in Eriksson and Dimitrakakis [9].

4.2 Decision-Making under Composite Risk Measures

A risk measure $U : \mathcal{X} \to \mathbb{R}$ is a function from a probability distribution to a scalar. This construction allows for decision-makers to compare risks under different distributions and choose what best adheres to their risk profile. One class of risk measures that has garnered a lot of interest recently is the *coherent* risk measures, given by Artzner *et al.* [24]. According to the definition, a coherent risk measure $U : \mathcal{X} \to \mathbb{R}$ has to satisfy four axioms:

Axiom 1 (Monotonicity)**:** If $X \leq Y$ almost surely, $U(X) \leq U(Y)$.

Axiom 2 (Positive homogeneity)**:** For any $c \geq 0$, $U(cX) = cU(X)$.

Axiom 3 (Translation invariance)**:** For any constant $a \in \mathbb{R}$, $U(X + a) = U(X) + a$.

Axiom 4 (Subadditivity)**:** For $X, Y \in \mathcal{X}$, $U(X + Y) \leq U(X) + U(Y)$.

In the work Eriksson *et al.* [10] our focus is on risk measures of this kind. One well-known coherent risk measure is CVaR and how it is impacted by varying the risk-sensitive parameter α for a few select distributions can be seen in Figure 4.2. The figure illustrates the decision-maker will significantly penalise behaviour that leads to high variability in the objective.

Quantifying Composite Risk Measures

Following Eriksson *et al.* [10] we define the risk measures of interest. To start with, we define the risk of the random variable Z under the distorted utility function U_α in three different ways for clarity.

$$\begin{aligned} \mathrm{Risk}_{U_\alpha}(Z) &\triangleq \int_{\mathcal{Z}} Z \,\mathrm{d}(U_\alpha \circ P) \\ &= \int_{\mathcal{Z}} U_\alpha(1 - F_Z(z)) \,\mathrm{d}z = \int_0^1 U_\alpha(t) \,\mathrm{d}q(1-t). \end{aligned} \tag{4.5}$$

Moving on with the risk measure associated with aleatory uncertainty, that is the uncertainty that arises due to the inherent stochasticity of the MDP μ and policy π, we chose the following definition.

Aleatory Risk. Given a coherent risk measure with distorted utility function U_α^A, the aleatory risk is quantified as the deviation of the total risk of

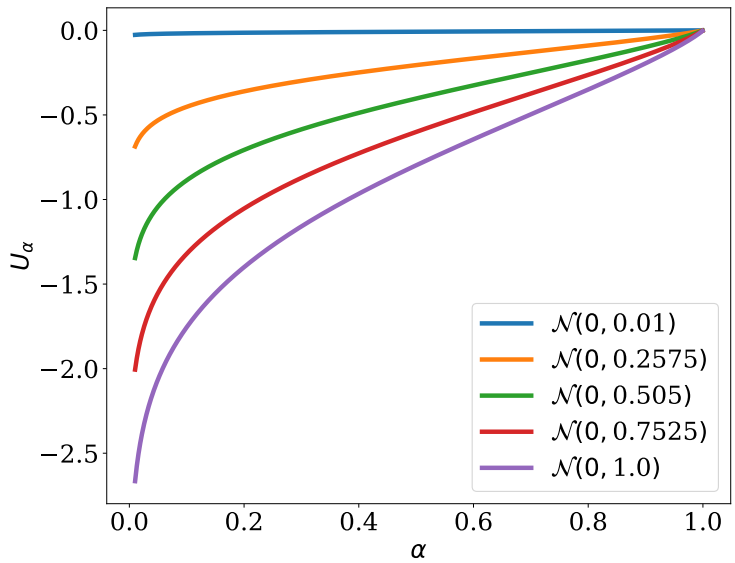


Figure 4.2: A small experiment demonstrating how the CVaR risk measure is impacted by changes in the risk-sensitive parameter α and by modifying the data distribution. Note that in this experiment only the left-tailed part of the distribution is considered and what is shown is $-\text{CVaR}_\alpha$.

individual models from the risk of the average model.

$$A(U_\alpha^A, \beta) \triangleq \int_\Theta \int_\mathcal{Z} Z \, \mathrm{d}(U_\alpha^A \circ \mathbb{P})(Z|\theta) \, \mathrm{d}\beta(\theta) - \int_\Theta \int_\mathcal{Z} \hat{Z} \, \mathrm{d}(U_\alpha^A \circ \mathbb{P})(\hat{Z})$$

Epistemic Risk. Given a coherent risk measure with distorted utility function U_α^E, the epistemic risk quantifies the uncertainty invoked by not knowing the true model. Thus, the risk can be computed over any statistics of the models, such as expectation.

$$E(U_\alpha^E, \beta) \triangleq \int_\Theta \int_\mathcal{Z} Z \, \mathrm{d}\mathbb{P}(Z|\theta) \, \mathrm{d}(U_\alpha^E \circ \beta)(\theta)$$

Composite Risk under Model and Inherent Uncertainty. Finally, in [10] a joint risk measure termed composite risk is defined that takes into account both the uncertainty that arises due to the true MDP μ being unknown, as well as the MDPs are inherently stochastic. The total uncertainty is then a combination of both these sources of uncertainty and in order to quantify the total uncertainty, we proposed *composite risk*.

Definition 3 (Composite Risk)**:** *For two coherent risk measures with distorted utility functions $U_{\alpha_1}^A$ and $U_{\alpha_2}^E$, belief distribution β on model parameters $\theta \in \Theta$, and a random variable $Z \in \mathcal{Z}$, the composite risk of epistemic and aleatory uncertainties is defined as*

$$\begin{aligned} F^C(U_{\alpha_1}^A, U_{\alpha_2}^E, \beta) &\triangleq \mathrm{Risk}_{U_{\alpha_2}^E}(\mathrm{Risk}_{U_{\alpha_1}^A}(Z|\theta)|\beta) \\ &= \int_\Theta \int_\mathcal{Z} Z \, \mathrm{d}(U_{\alpha_1}^A \circ \mathbb{P})(Z|\theta) \, \mathrm{d}(U_{\alpha_2}^E \circ \beta)(\theta) \\ &= \int_0^1 \int_0^1 U_{\alpha_2}^E(v) U_{\alpha_1}^A(u) \, \mathrm{d}q_{Z|\theta}(1-u) \, \mathrm{d}q_\beta(1-v) \end{aligned} \tag{4.6}$$

The inclusion of a composite risk measure allows for a more accurate representation of the total uncertainty compared to existing works optimising jointly over both risks, such as in [7], [20].

Theorem 5 (Coherence)**:** *If $U_{\alpha_1}^A$ and $U_{\alpha_2}^E$ are distorted utilities for two coherent risk measures, the composite risk measure $F^C(U_{\alpha_1}^A, U_{\alpha_2}^E, \beta)$ is also coherent.*

The theorem Theorem 5 is important so as to retain coherency after composing the risk measures.

Theorem 6: *We are given two sources of aleatory and epistemic uncertainties ξ_1 and ξ_2. If $U^A_{\alpha_1}$ and $U^E_{\alpha_2}$ are distortion measures for two coherent risk measures quantifying aleatory and epistemic risks respectively, then, i) $F^A(U^A_{\alpha_1}, \beta) = F^C(U^A_{\alpha_1}, I, \beta)$, where I is the identity function, and ii) $F^C(U^A_{\alpha_1}, U^E_{\alpha_2}, \beta) \geq F^A(U^A_{\alpha_1}, \beta)$, if $\alpha_2 \neq 1$.*

This theorem is used in the work Eriksson *et al.* [10] to demonstrate the superiority of the composed risk measure approach to an additive risk approach to jointly optimising for both risks. The proofs of the theorems Theorem 5 and Theorem 6 are left for the interested reader in the paper Eriksson *et al.* [10].

Optimising for Coherent Risk Measures

In our work, we propose an algorithm for optimising composite risk measures as defined in Eq. 4.6. The full algorithm is available in Algorithm 3.

Experimental Results using Coherent Risk Measures

The experimental results can be seen in the paper Eriksson *et al.* [10].

Algorithm 3 SENTINEL-K with Composite Risk

Input: Initial state s_0, action set $\mathcal{A}$, distortion measures $U^A_{\alpha_1}, U^E_{\alpha_2}$, hyperparameter λ, target networks $[\theta_1^-, ..., \theta_K^-]$, value networks $[\theta_1, ..., \theta_K]$, update schedule Γ_1, Γ_2.
for $t = 1, 2, \ldots$ **do**
  //* Update K-value and target networks for estimating return distributions *//
  for $t' \in \Gamma_1 \cup \Gamma_2$ **do**
    Generate $\{D_1, ..., D_K\} \leftarrow \text{DataMask}(\mathcal{D}^{t'})$
    for $i = 1, \ldots, K$ **do**
      Sample mini batch $\tau \sim D_i$
      $F^C(Z(s_t, a)|U^A_{\alpha_1}, U^E_{\alpha_2}, \beta)$ using τ and K-target networks $\{\theta_i^-\}_{i=1}^K$.
      Get $a^* = \arg\max_a F^C(Z(s_t, a)|U^A_{\alpha_1}, U^E_{\alpha_2}, \beta)$
      Update value network θ_i using τ, a^*
      Update target network θ_i^- using τ, a^* if $t' \in \Gamma_1$
    end for
  end for
  //* Estimate the composite risk of each action using the estimated return distributions *//
  for $a \in \mathcal{A}$ **do**
    Compute weights $\mathbf{w} = w_1, ..., w_K$.
    for i **in** K **do**
      Compute aleatory risks $Q_i^A(s_t, a)$ from $\int_{\mathcal{Z}} Z \, \mathrm{d}(U^A_{\alpha_1} \circ \mathbb{P})(Z|\theta_i)$
    end for
    Compute composite risk over weighted aleatory estimates $Q^C(s_t, a) = \text{Risk}_{U^E_{\alpha_2}}(\{w_i Q_i^A(s_t, a)\}_{i=1}^K)$
  end for
  //* Action selection *//
  Take action $a_t = \arg\max_a Q^C(s_t, a)$
  Observe s_t and update the dataset $\mathcal{D}^t \leftarrow \mathcal{D}^{t-1} \cup \{s_t, a_{t-1}, s_{t-1}, r_{t-1}\}$
end for

CHAPTER 5

Concluding Remarks and Future Work

In this chapter we conclude and tie together the discussions in Chapter 3, and Chapter 4. We also discuss future and ongoing work relating to this thesis.

5.1 Conclusion

In Chapter 3 we have presented a few ways of representing uncertainties in RL that we have worked on, namely work in Eriksson *et al.* [10] and Jorge *et al.* [12]. Construction of rich value-function representations allows then for risk-sensitive decision-making following the works in Eriksson and Dimitrakakis [9] and Eriksson *et al.* [10] discussed in Chapter 4.

5.2 Future Work

In a recent work in Eriksson *et al.* [25] we have extended the formulation in Eriksson and Dimitrakakis [9] to the Bayesian games setting. This setting involves multiple agents interacting simultaneously in the same environment. The uncertainty that arises due to the uncertainty in a Bayesian games is

similar to epistemic uncertainty in RL. The work proposes a joint optimisation routine for all agents, taking their individual utility functions and risk appetites into account. We also see a clear possibility of making a continuation of the work in Jorge *et al.* [12], adopting a risk-sensitive framework and extending the formulation from a backward induction one to one based upon optimisation using gradients.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[2] D. Silver, A. Huang, C. J. Maddison, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[3] D. Silver, J. Schrittwieser, K. Simonyan, *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[4] B. R. Kiran, I. Sobh, V. Talpaert, *et al.*, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[5] E. Leurent, "Safe and efficient reinforcement learning for behavioural planning in autonomous driving," Ph.D. dissertation, Université de Lille, 2020.

[6] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[7] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning," in *International Conference on Machine Learning*, 2018, pp. 1192–1201.

[8] T. Pearce, M. Zaki, A. Brintrup, N. Anastassacos, and A. Neely, "Uncertainty in neural networks: Bayesian ensembling," *stat*, vol. 1050, p. 12, 2018.

[9] H. Eriksson and C. Dimitrakakis, "Epistemic risk-sensitive reinforcement learning," in *ESANN*, 2020, pp. 339–344.

[10] H. Eriksson, D. Basu, M. Alibeigi, and C. Dimitrakakis, "Sentinel: Taming uncertainty with ensemble-based distributional reinforcement learning," *arXiv preprint arXiv:2102.11075*, 2021.

[11] O. Mihatsch and R. Neuneier, "Risk-sensitive reinforcement learning," *Machine learning*, vol. 49, no. 2-3, pp. 267–290, 2002.

[12] E. Jorge, H. Eriksson, C. Dimitrakakis, D. Basu, and D. Grover, "Inferential induction: A novel framework for bayesian reinforcement learning," 2020.

[13] D. P. Bertsekas and J. N. Tsitsiklis, "Neuro-dynamic programming: An overview," in *Proceedings of 1995 34th IEEE conference on decision and control*, IEEE, vol. 1, 1995, pp. 560–564.

[14] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 449–458.

[15] R. Dearden, N. Friedman, and S. Russell, "Bayesian q-learning," in *Aaai/iaai*, 1998, pp. 761–768.

[16] Y. Engel, S. Mannor, and R. Meir, "Reinforcement learning with gaussian processes," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 201–208.

[17] M. Hessel, J. Modayil, H. Van Hasselt, *et al.*, "Rainbow: Combining improvements in deep reinforcement learning," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[18] R. T. Rockafellar, S. Uryasev, *et al.*, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000.

[19] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," in *Advances in neural information processing systems*, 2016, pp. 4026–4034.

[20] W. R. Clements, B.-M. Robaglia, B. Van Delft, R. B. Slaoui, and S. Toth, "Estimating risk and uncertainty in deep reinforcement learning," *arXiv preprint arXiv:1905.09638*, 2019.

[21] K. Murphy, "Conjugate bayesian analysis of the gaussian distribution," UBC, Tech. Rep., 2007.

[22] C. Dimitrakakis, "Robust bayesian reinforcement learning through tight lower bounds," in *European Workshop on Reinforcement Learning (EWRL 2011)*, 2011, pp. 177–188.

[23] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

[24] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[25] H. Eriksson, D. Basu, M. Alibeigi, and C. Dimitrakakis, "Risk-sensitive bayesian games for multi-agent reinforcement learning under policy uncertainty," *arXiv preprint arXiv:2203.10045*, 2022.