博士論文

Multi-Stage Robust Decision Making: Decision Support Framework Under Deep Uncertainty and Its Application to Technology Roadmapping

（多段階ロバスト意思決定：深い不確実性の下での意思決定支援フレームワークと技術ロードマッピングへの応用）

by

Shunichiro Nomura

野村 俊一郎

A dissertation submitted to
Department of Aeronautics and Astronautics
School of Engineering
in partial fulfillemnt of the requirements for the degree of

Doctor of Philosophy

at

The University of Tokyo

Abstract

In large and complex projects, it is crucial to acknowledge uncertainties and make decisions that perform well in a wide range of future scenarios. This is because the point estimate of the future is often inaccurate, and even if we know the future uncertainty accurately, making plans only based on the most likely future may results in a bad outcome due to the flaw of averages. It is also worth considering that those uncertainties are often deep, i.e., non-probabilistic, making it challenging to apply conventional probabilistic analyses.

Multi-stage, or sequential, decision making is often effective under a deeply uncertain environment. This dissertation proposes a multi-stage-robust-decision-making Markov decision process (MSRDM-MDP), an extension of a Markov decision process, that can model sequential decision making under deep uncertainty. We show that the maximax and maximin optimal policies can be obtained by solving the maximax and maximin optimal Bellman equations using a reinforcement learning algorithm.

This dissertation also proposes a horizon-of-uncertainty (HoU) analysis that helps decision-makers understand the trade-off between each policy option's performance and robustness.

Based on these proposed concepts, this dissertation proposes a computer-aided decision-support framework called multi-stage robust decision making (MSRDM) that helps decision-makers make better decisions even under non-probabilistic uncertainties by enabling them to frame the problem as a multi-stage decision-making problem and analyze the trade-off between the performance and the robustness of each policy option.

Finally, the proposed framework is demonstrated in two case studies: technology roadmapping of the space formation flying system and technology roadmapping of the marine propulsion system.

Acknowledgements

It would not have been possible to complete my three-year Ph.D. and write this Ph.D. dissertation without the help and support of the kind people around me.

First and foremost I am extremely grateful to my supervisor, Shinichi Nakasuka, for his invaluable advice, and continuous support not only during my Ph.D. study, but also my entire six years in his laboratory. His immense knowledge and experience always helped me grasp a wider, practical perspective on issues.

I would also like to thank my thesis committee members. Koichi Hori suggested to me positioning my research with respect to other research domains in decision-making such as creativity support systems, which clarified my research scope. Akira Iwasaki, an expert in the remote sensing technology, informed me with needs and issues in actual decision-making situations from practical perspectives. Ryu Funase has been regularly giving advice, support, and irreplaceable experiences in my years in the laboratory, which cannot be covered just by this dissertation. Kazuo Hiekata has provided me, who is unfamiliar with the maritime field, with valuable feedback on the maritime case study. He also gave me an opportunity to visit MIT SDM and discuss research topics with researchers with the same research interest, which did, and surely will, widen my research perspectives.

I am also grateful to those who offered me interesting and practical case studies: Satoshi Ikari, Yasuo Ichinose, and Shinnosuke Wanaka. Without the case studies it would be difficult to demonstrate the proposed framework's values.

I would also like to acknowledge the insightful feedback and suggestions from Bryan Moser, Hiroshi Sasaki, Takuto Ishimatsu, Koki Ho, Marc-Andre Chavy-Macdonald,

Tomoki Omiya, Satoshi Hirayama, Shun Furuya, Ryo Matsuoka, and the laboratory members, especially those who belong to the SE/AI research group.

I must also thank Masahiro Ono and Kyohei Otsu for letting me join their Mars rover research project and giving me an opportunity to learn the basic of robotics and discuss my research with experts at JPL.

I also wish to thank the Global Leader Program for Social Design and Management (GSDM) for various kinds of opportunities to gain new perspectives as well as the financial support.

Finally, and importantly, the completion of my dissertation would not have been possible without the support and nurturing of my parents and my sisters. Ultimately, I would like to thank my partner Maho Nomura, who brings the joy of life to my day and is my daily support.

Table of contents

List of figures

List of tables

Nomenclature

Roman Symbols

$\mathcal{A}$	set of all actions
a	action
$\mathcal{B}$	set of all beliefs
b	belief
C_0	cumulative reward
c_i	development cost of technology i
CVaR	conditional value at risk, or expected shortfall
Δw_i	maximum deviation of w_i from the nominal value
d_i	distance function of uncertain parameter w_i
d_{w}	number of uncertain parameters, i.e., dimension of scenario w
$\mathbb{E}[\cdot]$	expectation operator
e	Euler's number (≈ 2.71828)
$\mathbb{E}^{\pi}$	expectation operator conditioned by policy π
f	objective function

$f_{\text{feas},j}$ mission feasibility function

$f^{-}_{\text{HoU}}(h; s_0, b_0, \pi)$ minimum performance measure in $\mathcal{U}(h)$

$f^{+}_{\text{HoU}}(h; s_0, b_0, \pi)$ maximum performance measure in $\mathcal{U}(h)$

$f_{\text{preq},i}$ development readiness function

$\mathcal{H}_t$ set of all histories

h horizon of uncertainty

h_t history

$\mathcal{L}$ set of lotteries

Lognormal(μ, σ^2) Log-normal distribution with mean μ and standard deviation σ

$\mathcal{M}$ Markov decision process

$\mathcal{M}(\pi)$ Markov decision process under policy π

m the number of configurations in the `MarinePropulsion` problem

m the number of missions in the `FormationFlying` problem

$\mathcal{M}_{\text{G}}$ Markov game, maximin Markov decision process

$\mathcal{M}_{\text{MSRDM}}$ multi-stage-robust-decision-making Markov decision process

$\mathcal{N}(\mu, \sigma^2)$ Normal distribution with mean μ and standard deviation σ

n the number of technologies

n_{ep} number of episodes completed

n_{ε} decay factor for ε scheduling in ε-greedy policy

n_{w} number of scenarios

p_j reward for mission j

p_{s_0} initial state probability function

Q^π action-value function

Q^* optimal action-value function

$\mathbb{R}$ set of all real numbers

R reward function

$R_{\max}$ the upper bound of reward function

r_{c} critical value

r_{w} windfall value

$\mathcal{S}$ set of all states

s state

T transition function

t current time

T_i development time of technology i

$t_{\lim}$ time limit

t_{rev} fuel scenario reveal

T_{s} state transition function

$\mathcal{U}$ non-probabilistic uncertainty model

U utility function

$U(a, b)$ the continuous uniform distribution in $[a, b]$

u_i	technology i's time under development
VaR	value at risk
v_i	Boolean flag indicating whether development of technology i is completed
V^π	value function
V^*	optimal value function
V_-^*	maximin optimal value function
V_+^*	maximax optimal value function
$\mathcal{W}$	set of all scenarios
W	random variable
w	scenario
w	uncertain parameter in Chapter 1
$w_{\mathrm{P}}^{(k)}$	Pareto scenario
$\tilde{w}$	nominal scenario
$w_{\mathrm{v}}^{(i)}$	Pareto vertex
x	decision variable
Y	value of the objective function as a random variable

Greek Symbols

$\hat{\alpha}$	info-gap robustness
β_i	budget-of-uncertainty function of uncertainty parameter w_i
$\hat{\beta}$	info-gap opportuneness

ε	probability of selecting a random action under ε-greedy policy
ε_{f}	limit of ε in ε-greedy policy
ε_{i}	initial value of ε in ε-greedy policy
η_j	CO_2 emission reduction performance of configuration j
γ	discount factor
μ	mean
ϕ	fuel scenario
Π	set of all deterministic Markov policies
π	deterministic Markov policy
Π^{HD}	set of all deterministic history-dependent policies
π^{h}	deterministic history-dependent policy
π^*	optimal policy
π^*_-	maximin optimal policy
π^*_+	maximax optimal policy
ρ	risk measure
σ	standard deviation
θ	neural network parameters

Superscripts

LB	lower bound
$*$	optimal

UB upper bound

Subscripts

t time step

Acronyms / Abbreviations

CDMSS creativity enhancing decision-making support system

CoI cases of interest

DEM digital elevation model

DLR German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt)

EEDI energy efficiency design index

FDIR fault detection, isolation, and recovery

GAO Government Accountability Office

GHG greenhouse gas

GRACE Gravity Recovery and Climate Experiment

HCSC The Health Care Service Corporation

HoU horizon of uncertainty

ICE Internal combustion engine

IMO International Maritime Organization

LNG liquefied natural gas

MDP Markov decision process

MSRDM-MDP Multi-stage-robust-decision-making Markov decision process

MSRDM Multi-stage robust decision making

NASA National Aeronautics and Space Administration

PRIM patient rule induction method

RDM Robust decision making

RL reinforcement learning

TanDEM-X TerraSAR-X add-on for Digital Elevation Measurement

Chapter 1

Introduction

1.1 Motivation

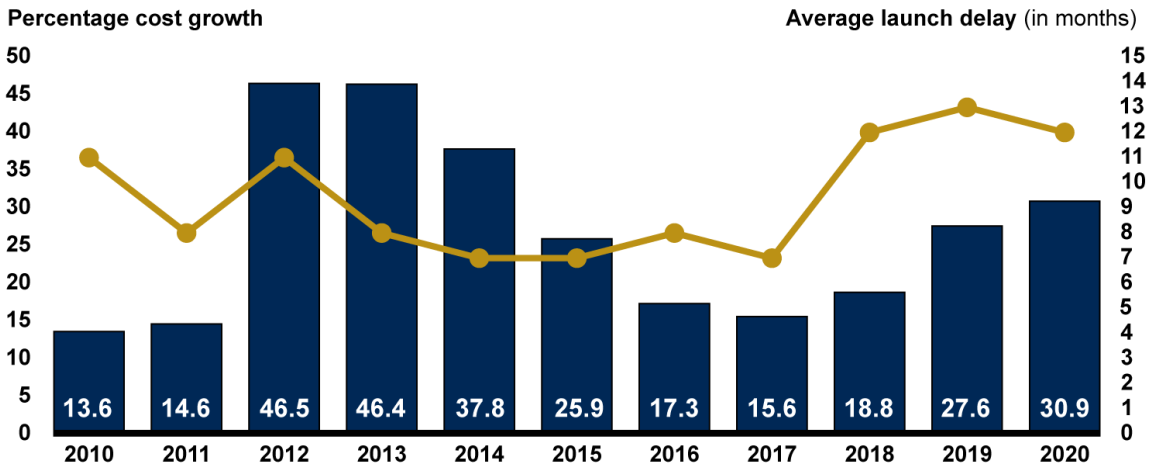
Designing and operating a complex system often involve large uncertainties. Take space exploration as an example. The uncertainties in space exploration can be categorized into at least three types: political, technical, and "pure" uncertainties.

Political uncertainty includes that of space policies of governments. Table 1.1 shows the U.S. human space flight program and its primary destination have changed each time a new administration takes office since President George W. Bush. This is considerable uncertainty for other governments and space agencies, whose space policies are often affected by the U.S. space policy. Another example of political uncertainty is the international legal frameworks on space debris and space resource utilization [1–8].

Table 1.1 NASA Human spaceflight programs and their primary destinations under different administrations.

Administration	Human spaceflight program	Primary destination
George W. Bush	The Constellation program [9, 10]	Moon
Barack Obama	Journey to Mars[1][11]	Asteroid [12–17], Mars [11]
Donald Trump	The Artemis program [18]	Moon

[1] Journey to Mars was technically not the name of NASA's human space program, but the name of the overall NASA's space exploration strategy.



Source: GAO analysis of National Aeronautics and Space Administration data. | GAO-20-405

Figure 1.1 Development cost performance and average launch delay for major NASA projects from fiscal year 2010 through fiscal year 2020 [19].

Primary sources of technical uncertainty include development cost, development time, and realized performance. The U.S. Government Accountability Office (GAO) [19] reported that major projects of the National Aeronautics and Space Administration (NASA) experienced cost growth of 31 percent over the project baselines and an average launch delay of 12 months, as shown in Figure 1.1. The Space Shuttle program famously experienced uncertainty. It was initially designed to fly routinely and make access to space more inexpensive. However, although the original plan was to fly the shuttle up to 60 times a year, the flight frequency was about eight flights per year [20]. The average cost per launch from 1991 to 2010 was $1.5 billion per launch in 2010 dollars [21], more than 27 times costlier than the original estimate of the cost per launch: $54.3 million[1] [22].

"Pure" uncertainty is due to a lack of knowledge about the environment of space. For example, direct evidence of surface-exposed water ice on the Moon has recently been reported [23], and some private companies aim to extract and sell it [24]. However, uncertainty in the lunar ice's characteristics (e.g., how abundant the water ice is, what the mining rate will be with each mining technology) makes the lunar water

[1] Converted from $10.4 million in 1972 dollars with the conversion rate: $1 in 1972 = $5.22 in 2010.

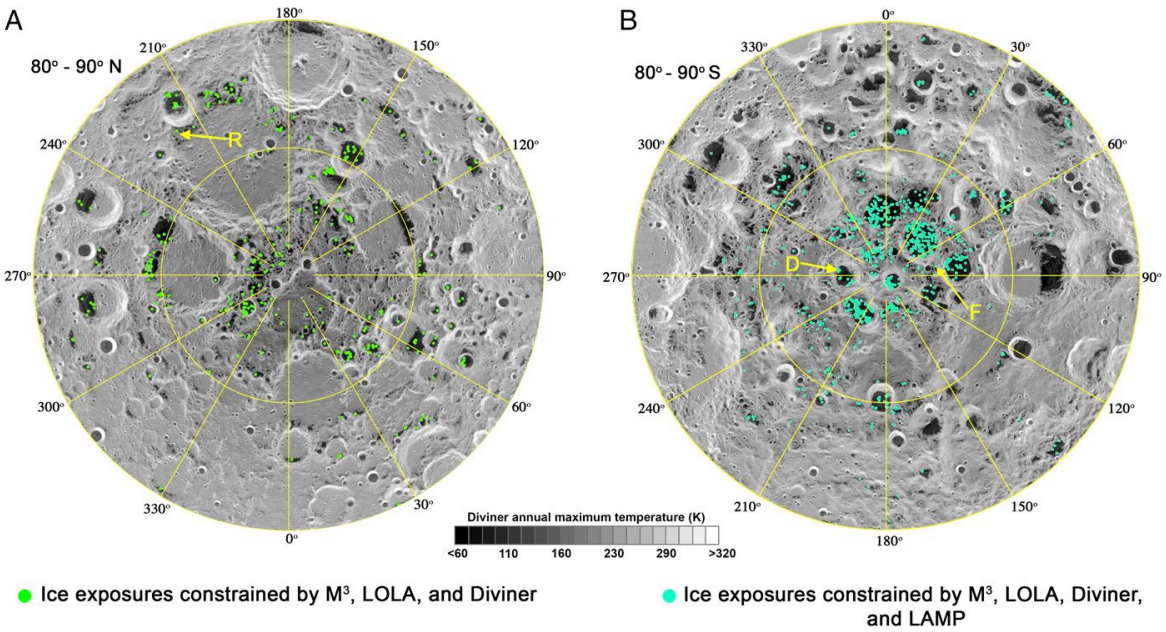


Figure 1.2 Distribution of water-ice-bearing pixels (green and cyan dots) overlain on the Diviner annual maximum temperature for the (A) northern- and (B) southern polar regions [23].

mining business, which requires an enormous amount of up-front investment, even more difficult.

Uncertainty also exists and has been studied in the context of climate change and CO_2 emission reduction technologies [25–28], petroleum exploration and production [29–33], and water resource management [34–38].

This research aims to support robust and adaptive decision making under uncertainty (in particular, *deep uncertainty*, described in Section 1.2.1) by proposing a decision-support framework for multi-stage decision making under uncertainty aided by reinforcement learning.

1.2 Background

1.2.1 Types of uncertainty

Before discussing how to make decisions under uncertainty, we need to understand uncertainty because different uncertainties have different properties and should be handled accordingly.

Aleatory uncertainty, epistemic uncertainty, and Talebian uncertainty

One of the categorizations of uncertainty is based on its source, categorizing uncertainties into three types: *aleatory uncertainty*, *epistemic uncertainty*, and *Talebian uncertainty* [39].

Aleatory uncertainty, also known as irreducible uncertainty or intrinsic uncertainty, is uncertainty due to inherent variability in a physical phenomenon. Let us consider a box with balls in it. If we know that half of the balls are red and the other half are white, then the color of a ball we pick randomly from the box will be red with 50 % probability or white with 50 % probability, as shown in Figure 1.3a. It is aleatory uncertainty because this uncertainty of the color is intrinsic in the box's physical state. As the alias "irreducible uncertainty" suggests, we cannot reduce the uncertainty.

Epistemic uncertainty, also known as reducible uncertainty or knowledge uncertainty, is uncertainty due to a lack of knowledge about the event. Suppose we know that the balls in the box are all red with 50 % probability, or all white with 50 % probability, as shown in Figure 1.3b, then the color of a ball we pick randomly from the box will be red with 50 % probability, or white with 50 % probability. It is epistemic uncertainty because it is not intrinsic in the physical state but the lack of knowledge about the balls' color. Although the color of the randomly-picked ball has the same probability distribution as the case of aleatory uncertainty, it is different in that we can reduce the uncertainty by observing the color of the balls in the box. Note that whether uncertainty is categorized as aleatory uncertainty or epistemic uncertainty is not always obvious. We could say, for example, that the uncertainty in the color of a

randomly-picked ball from the box with the same number of red balls and white balls (Figure 1.3a) is not aleatory but epistemic because the uncertainty derives from the fact that we do not know the location and the color of each ball in the box. If the box is transparent, we may be able to pick a red ball without uncertainty deliberately.

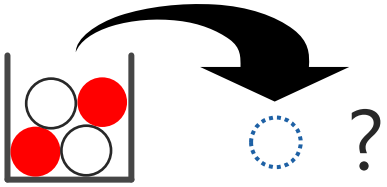
Talebian uncertainty, also known as ignorance, model uncertainty, or unknown unknowns, is an uncertainty that is not considered or cannot be anticipated because the prior model of the event is wrong. Suppose we think that the balls in the box are all red or white, but the randomly-picked ball turns out to be black. The uncertainty derives from the lack of knowledge about the event, and it, by definition, is impossible to anticipate before it occurs. The name "Talebian" comes from *Nassim Taleb*, who proposed the concept of Black Swan, uncertainty with a major effect that comes as a surprise [40].

Risk and true uncertainty

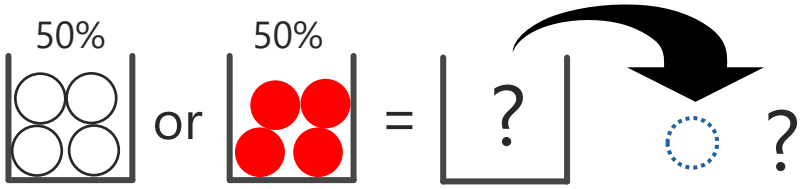
Another categorization is based on whether its probability distribution is known, and it categorized uncertainties into two types: *risk* and *true uncertainty* [41, 42], visualized in Figure 1.4.

Risk, or probabilistic uncertainty, is uncertainty whose probability distribution is known. The examples in Figures 1.3a and 1.3b are both probabilistic uncertainty because we know the probability distribution of the color of a randomly-picked ball, that is, 50 % red and 50 % white. Once the probability distributions of all uncertainties are known, we can apply probabilistic calculations to obtain the probability distribution of outcomes and make decisions based on the analysis results.

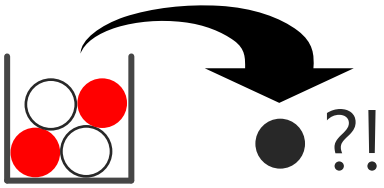
However, the probability distributions of uncertain parameters are not always available to the decision-makers or even experts. This uncertainty is called true uncertainty, Knightian uncertainty, ambiguity, or non-probabilistic uncertainty. If there is a box and we have no information on what is contained in it, what a randomly-picked object will be is true uncertainty because we know so little that we do not even know the probability distribution of what the object will be.



(a) Aleatory uncertainty

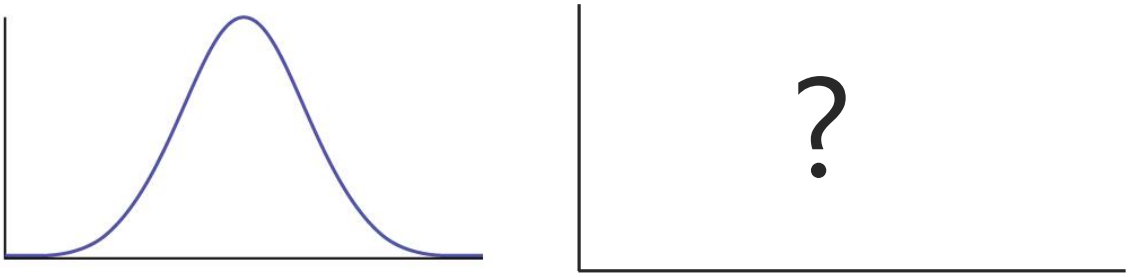


(b) Epistemic uncertainty



(c) Talebian uncertainty

Figure 1.3 Three types of uncertainty categorized based on its source.



(a) Risk

(b) True uncertainty

Figure 1.4 Two types of uncertainty categorized based on whether its probability distribution is known: *risk*, whose probability distribution is known, and *true uncertainty*, whose probability distribution is unknown.

There is a more generalized concept called *deep uncertainty* [43, 44], which R. J. Lempert et al. defined in [43] as:

> Deep uncertainty exists when analysts do not know, or the parties to a decision cannot agree on, (1) the appropriate models to describe the interactions among a system's variables, (2) the probability distributions to represent uncertainty about key variables and parameters in the models, and/or (3) how to value the desirability of alternative outcomes.

When facing deep uncertainty, applying probabilistic analysis is difficult or, if possible, not valuable. The following list shows examples of deep uncertainty, categorized into discoveries, inventions & innovations, and surprised. The categorization and some of the examples (marked with *) are from Ben-Haim, Y. [45].

- Discoveries
  - Discovery of the American continent (to people in Europe)*
  - Nuclear fission*
  - Discovery of life on Mars*
  - Discovery of water on the Moon
- Inventions & Innovations
  - Printing press (material invention)*
  - Ecological responsibility (conceptual innovation)*
  - French revolution (social innovation)*
  - Reusable rockets
  - Space traffic management
- Surprises
  - Competitor's innovation*

- Natural catastrophe*
- The Kessler syndrome[2]

1.2.2 Why should we consider uncertainty in decision making?

As discussed in Section 1.2.1, there are various types of uncertainties that may affect the outcome of interest of a decision-maker. It is often essential for a decision-maker to recognize uncertainty in the problem and reflect it in the decision for two reasons: *prediction difficulty* and *the flaw of averages* [47, 48].

Prediction difficulty is the difficulty in estimating the true value of an uncertain parameter. In a complex system, especially under deep uncertainty, estimating not the range but the value of an uncertain parameter (this is called "point estimate") is difficult because of the limitation in observability and understanding of the dynamics behind it. An extreme example is McKinsey & Co.'s prediction in 1980 of the number of mobile phone users in the U.S. in 2000. Their estimate was 900,000 [49], which turned out to be less than 1 % of the actual value: 109 million [50, 51].

Suppose we have an estimate of the probability distribution of uncertain parameters. In that case, we should explicitly consider the various realizations of the uncertain parameters in the decision-making process. Using a single representative value such as the expected value as the point estimate may yield an unwanted outcome, even if the knowledge of the distribution is accurate. This is called the flaw of averages. According to Sam Savage [48], the proposer of the concept, it states that "(p)lans based on *average* assumptions are wrong on *average*." He gave an example of a "statistician who drowned while fording a river that was, on average, only three feet deep." See Figure 1.5 for the illustration by Jeff Danziger. To show this property, let us assume a problem with a decision variable $x \in \mathbb{R}$, an uncertain parameter $w \sim \mathcal{N}(0, 1)$, and an

[2]The total amount of space debris will increase by itself once it surpasses a certain threshold because a collision leads to more debris, leading to more collisions, in a chain reaction. This is called the Kessler syndrome [46].

objective function $f(x, w) = (x - \mathrm{e}^w)^2$ to minimize. If the decision-maker knows the probability distribution of w but optimizes x only for the expected value of w, then the optimal decision x_1^* will be as follows:

$$\begin{aligned} x_1^* &= \underset{x}{\operatorname{argmin}}\, f(x, \mathbb{E}[w]) \\ &= \underset{x}{\operatorname{argmin}}\, f(x, 0) \\ &= \underset{x}{\operatorname{argmin}} (x-1)^2 \\ &= 1 \end{aligned} \tag{1.1}$$

where $\mathbb{E}[\cdot]$ denotes the expected value. However, if the decision-maker optimizes x for the distribution of w by minimizing the expected value of $f(x, w)$, the optimal decision x_2^* will be as follows:

$$\begin{aligned} x_2^* &= \underset{x}{\operatorname{argmin}}\, \mathbb{E}\left[f(x, w)\right] \\ &= \underset{x}{\operatorname{argmin}}\, \mathbb{E}\left[(x - \mathrm{e}^w)^2\right] \\ &= \underset{x}{\operatorname{argmin}} \left(x^2 - 2\mathbb{E}\left[\mathrm{e}^w\right] x + \mathbb{E}\left[\mathrm{e}^{2w}\right]\right) \\ &= \underset{x}{\operatorname{argmin}} \left(x^2 - 2\mathrm{e}^{\frac{1}{2}} x + \mathrm{e}^2\right) \\ &= \underset{x}{\operatorname{argmin}} \left[\left(x - \mathrm{e}^{\frac{1}{2}}\right)^2 + \mathrm{e}(\mathrm{e}-1)\right] \\ &= \mathrm{e}^{\frac{1}{2}} \approx 1.649 \end{aligned} \tag{1.2}$$

Note that e^w and e^{2w} are both log-normally distributed ($\mathrm{e}^w \sim \mathrm{Lognormal}(0, 1), \mathrm{e}^{2w} \sim \mathrm{Lognormal}(0, 2^2)$), and the mean of a log-normal distribution $\mathrm{Lognormal}(\mu, \sigma^2)$ is $\exp\left(\mu + \frac{\sigma^2}{2}\right)$. The difference between x_1^* and x_2^* comes from the fact that the asymmetricity in e^w with respect to the mean of the uncertain parameter $w = 0$ is not considered in solving x_1^*. More specifically, because the deviation of e^w when $w > 0$ is larger than that of e^w when $w < 0$, the decision-maker should pay more attention to the uncertainty of $w > 0$ by increasing x from x_1^*. This corresponds to the uncertainty that has a small probability but has a major effect.

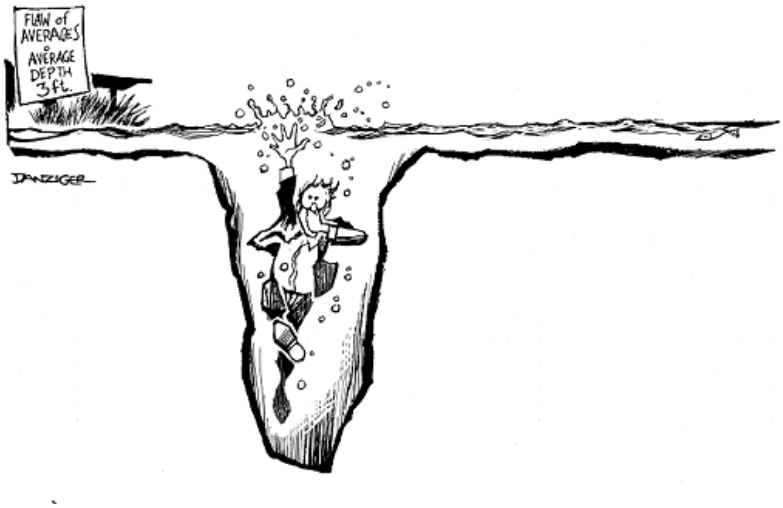


Figure 1.5 A "statistician who drowned while fording a river that was, on average, only three feet deep." An illustration in [48] by cartoonist Jeff Danziger.

As discussed in [52] as the "architecture robustness principle," a decision-maker should take into account three properties in selecting a complex system architecture: *optimality*, *robustness* (capability of dealing with environment changes without transforming itself), and *adaptability* (capability of transforming itself and adapting to environment changes)[3]. As [52] states that the optimal architecture is often the one with least robustness, it is necessary to consider the trade-off between the performance (optimality) and the robustness of design options.

1.2.3 Multi-stage decision making

When making decisions under uncertainty, it is often a good strategy to consider the time axis and make the decision adaptive, i.e., let the decision change flexibly based on new information available to the decision-maker during the operation. Adaptive strategies include the wait-and-see strategies, information-gathering actions, and a strategy to prepare for a highly-rewarding scenario with low probability.

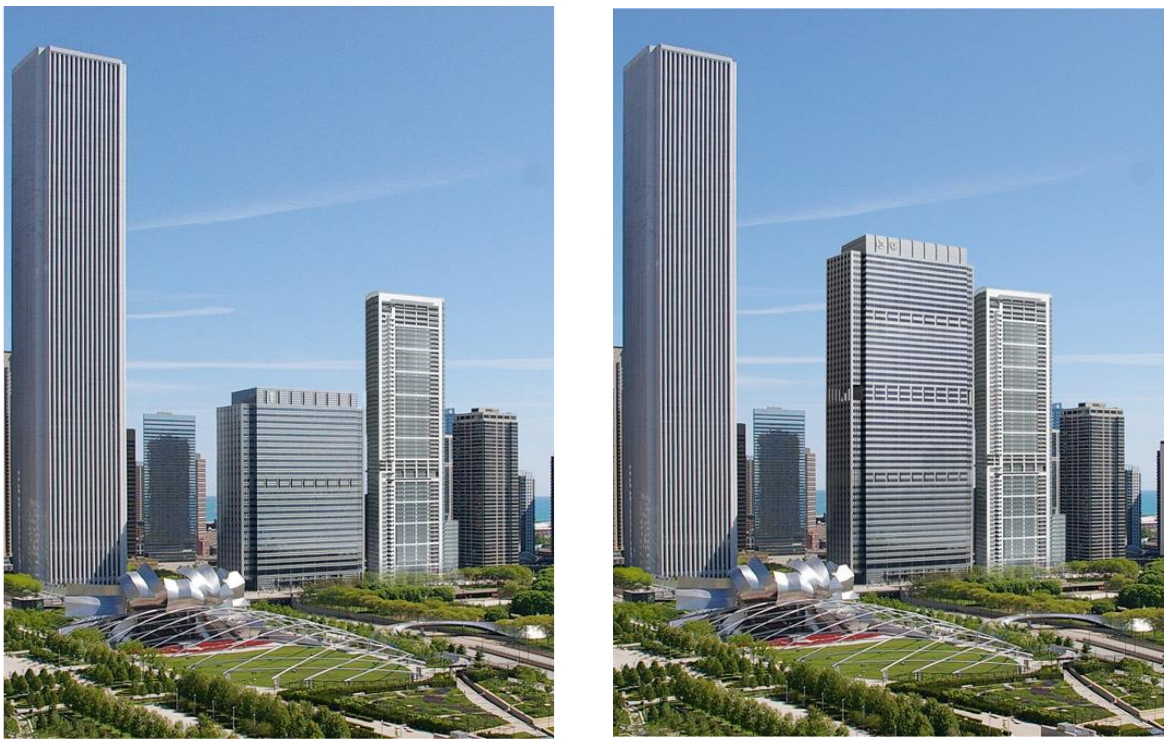
[3]Robustness and adaptability in [52] are different but both referred as "robustness" in this dissertation.

The wait-and-see strategy is a strategy to wait until the uncertainty of interest is reduced enough for the decision-maker to make a major decision. Such uncertainties include standardizations in industries and the development of infrastructure or core technology needed for a new product. For example, a car company can wait and see if a sustainable supply of Lithium-ion batteries is established or if the hydrogen fuel supply infrastructure is established before they decide to invest in electric cars powered by Lithium-ion batteries or ones powered by hydrogen fuel cells. If we adopt this strategy, we need to know the parameters to monitor and the conditions under which we should stop waiting and act.

The information-gathering action is an action that is taken only to collect information. It is of no value if we ignore the time axis because it does not generate any value at the time of the action. A similar concept called "active sensing" can be found in the control theory or behavioral science [53], which controls the system to gather information on the system's state and environment. It is important to understand what uncertain parameter has high sensitivity and is worth investigating because information-gathering often requires cost, both in money and time. It can be distinguished from the wait-and-see strategy in that it is active information gathering while the other is passive.

A strategy to prepare for a highly-rewarding scenario with low probability can be found in many start-ups and R&D projects. Even if the probability of a scenario under which the decision-maker can receive a large amount of reward is low, they can invest in projects necessary to enjoy the benefit under the rare scenario. Once the uncertainty is reduced and whether the scenario is realized becomes clear, they can increase the investment in the projects if the scenario is realized or terminate the projects otherwise.

Multi-stage, flexible decision-making has been investigated in real estate development [54].Vertical phasing is an example of multi-stage, flexible strategy. Vertical phasing is described in [55] as "constructing first a shorter building and then adding significant expansion later by increasing the building's height." The Health Care Service Corporation (HCSC) headquarters building in downtown Chicago was constructed



(a) Initial phase (b) Vertical completion phase

Figure 1.6 The Health Care Service Corporation building in the initial and vertical completion phases [57, 58, 55]. Source: Goettsch Partners, 2008.

in two phases. In phase 1, it was constructed in 1997 as a 33-story building that provides 1,430,000 square feet of space, and the foundations and structure were planned, designed, and constructed to support the fully expanded building. In phase 2, 24 additional stories were built on top of the existing building to provide additional 920,000 square feet of space [56]. HCSC planned for vertical phasing of their headquarters because they "did not want to commit to what it might need in the 2010's and beyond [55]."

A staged, flexible strategy is also considered in the communication satellite constellations deployment [59]. They proposed a flexible constellation deployment strategy where the constellation is progressively deployed and reconfigured according to the unfolded demand and showed the benefits of the staged approach when facing large demand uncertainty.

All adaptive strategies shown above can be valuable in decision making under uncertainty. Decision-makers need to make decisions in multiple stages in the time axis and should not stick to a single, fixed plan made at the beginning.

1.3 Thesis contributions

This research:

1. proposes the multi-stage-robust-decision-making Markov decision process (MSRDM-MDP), an extension of the Markov decision process (MDP) that can model multi-stage decision making under deep uncertainty and can be solved using the reinforcement learning.

2. proposes the multi-stage robust decision making (MSRDM), a quantitative human-in-the-loop decision-support framework using the MSRDM-MDP and the reinforcement learning for decision making under deep uncertainty.

3. validates the effectiveness of the MSRDM by applying it to two case studies: technology roadmapping in the space formation flight system and technology roadmapping in the marine propulsion system.

Table 1.2 compares our proposed framework with existing decision-support frameworks under uncertainty. Note that the proposed framework handles aleatory uncertainty and epistemic uncertainty, but not Talebian uncertainty (unknown unknowns) because the framework requires the decision-maker to be aware of the uncertainty.

1.4 Thesis structure

The remainder of this dissertation is structured as follows: Chapter 2 reviews the literature in decision making under probabilistic uncertainty, decision making under deep uncertainty, uncertainty management in practice, creativity support systems, and technology roadmapping methods. Chapter 3 presents the definition of the

Table 1.2 Comparison with exiting literature.

	Handles deep uncertainty	Provides robustness–performance trade-off	Models multi-stage decision making
Robust decision making [43, 60]	✓		
Engineering options analysis [54]			✓
Info-gap decision theory [61]	✓		
Markov decision processes			✓
Risk-aware Markov decision processes [62]		✓	✓
Multi-stage robust decision making (ours)	✓	✓	✓

MSRDM-MDP, its Bellman equation, and a reinforcement learning algorithm to solve the optimal policy. Chapter 4 presents the MSRDM framework and describes each step in the process with a toy problem, `SimpleMining`. Chapter 5 and Chapter 6 present the results of applying the proposed framework to two case studies: technology roadmapping of the space formation flight system and the marine propulsion system to show the proposed framework's effectiveness. Finally, Chapter 7 summarizes the proposed framework and the results of the two case studies and discusses the potential of future research.

Chapter 2

Literature Review

2.1 Decision-making under probabilistic uncertainty

2.1.1 Risk measures

Let us define a problem with a decision variable $x \in \mathcal{X}$, an uncertain parameter $w \in \mathcal{W}$, and an objective function $f(x, w)\colon \mathcal{X} \times \mathcal{W} \to \mathbb{R}$ to maximize. Let us consider a random variable $W \in \mathcal{W}$ and a decision x, then the outcome of the objective function is also a random variable: $Y = f(x, W)$. A risk measure ρ is a mapping from a random variable to a scalar. In the case of this problem, the risk measure $\rho(Y)$ represents the value of a decision x under a random variable W.

A common risk measure is the expected value of the objective function

$$\text{The expected value} \equiv \mathbb{E}\left[Y\right] \tag{2.1}$$

where $\mathbb{E}[\cdot]$ is the expectation operator. It is risk-neutral in that it is not affected by the degree of uncertainty.

Another risk measure is the value at risk, or α-quantile, defined as:

$$\text{VaR}_\alpha(Y) \equiv \inf\left\{y \in \mathbb{R} \mid \ \Pr(Y < y) > \alpha\right\} \tag{2.2}$$

It is the minimum value that the probability of Y less than the value is greater than α. Generally, the probability of the objective function being less than the value at risk is α[1], i.e., $\Pr(Y < \mathrm{VaR}_\alpha(Y)) = \alpha$.

The expected shortfall, or the conditional value at risk [63], is the objective function's expected value in the worst cases. Let α the quantile, then the conditional value at risk is calculated as:

$$\mathrm{CVaR}_\alpha(Y) \equiv \frac{1}{1-\alpha} \int_{-\infty}^{\mathrm{VaR}_\alpha(Y)} yp(y)\mathrm{d}y \tag{2.3}$$

where $p(y)$ is the probability density function of Y.

Let $\mathcal{L}$ be a set of lotteries and consider lotteries $L, M, N \in \mathcal{L}$. A lottery is a possibly random alternative. The random variable Y under a fixed decision x is an example of a lottery. When the decision-maker is indifferent between L and M, we write $L \sim M$, and we write $L \succeq M$ when the decision-maker prefers L over M or is indifferent. The *utility function* $U\colon \mathcal{L} \to \mathbb{R}$ assigns a value to lotteries based on the decision-maker's preference so that if the decision-maker prefers a lottery to another, then its utility is higher than the other. Formally, $\forall L, M \in \mathcal{L}\colon L \succeq M \implies U(L) \geq U(M)$. The expected utility is the expected value of the utility:

$$\text{The expected utility} \equiv \mathbb{E}[U(Y)] \tag{2.4}$$

It is proven that under the following four axioms, a decision-maker will act to maximize the expected value of a function, known as the von Neumann-Morgenstern utility function. The four axioms are:

Completeness $\forall L, M \in \mathcal{L}\colon L \succeq M \vee M \succeq L$

Transitivity $\forall L, M, N \in \mathcal{L}\colon$ if $L \succeq M$ and $M \succeq N$, then $L \succeq N$

Continuity if $L \succeq M \succeq N$, then $\exists p \in [0,1]$ s.t. $pL + (1-p)N \sim M$

[1] This is not always the case if Y is a discrete random variable.

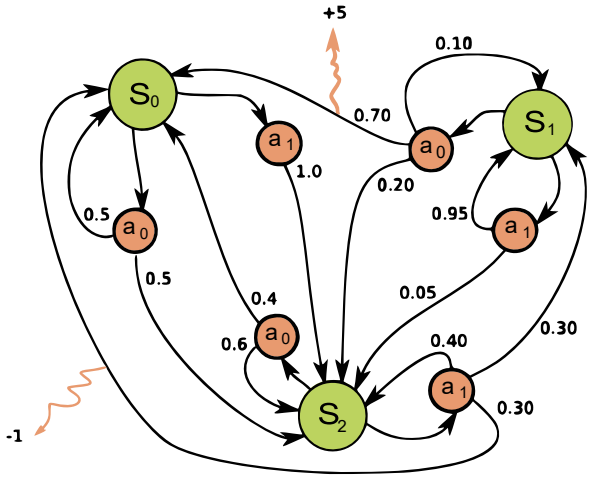


Figure 2.1 An example of a Markov decision process

Independence $\forall L, M, N \in \mathcal{L}, p \in [0,1]$: if $L \succeq M$, then $pL + (1-p)N \succeq pM + (1-p)N$

The expected value is consistent with the axioms. Yamai and Yoshiba [64] showed that the expected shortfall is also consistent, and the value-at-risk is consistent only under some conditions.

The above risk measures require the probability density function of Y to be known. However, if it is not the case, i.e., the uncertainty is "deep," the maxi-minimality is available, if technically not a risk measure. The maxi-minimality is to select the decision with the best worst-case scenario by solving the following optimization problem:

$$\max_{x \in \mathcal{X}} \min_{w \in \mathcal{W}} f(x, w) \tag{2.5}$$

This formulation is studied as "robust optimization," and the decision-maker's attitude towards risk can be represented by the uncertainty set $\mathcal{W}$ [65, 66].

2.1.2 Markov decision processes

A Markov decision process (MDP) [67] is a model of an environment where an agent stochastically transits from a state to another based on its action while receiving a

reward. If a probabilistic model of the environment is provided as an MDP, the optimal policy to maximize the expected cumulative reward can be solved.

An MDP $\mathcal{M}$ can be formally defined as a tuple of five elements [68]:

$$\mathcal{M} \equiv \langle \mathcal{S}, \mathcal{A}, p_{s_0}, T, R \rangle \tag{2.6}$$

where $\mathcal{S}$ is a set of all states, $\mathcal{A}$ is a set of all actions, $p_{s_0} \colon \mathcal{S} \to [0, 1]$ is the initial state probability function, $T \colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function, and $R \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function. Here we consider a discrete state set and a discrete action set. Therefore, by definition, the reward function R is upper-bounded, and there exists $R_{\max}$ that satisfies:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} \colon |R(s, a)| \leq R_{\max} \tag{2.7}$$

The decision variable in an MDP is called *policy*. A policy can be formulated in various ways, one of which is a deterministic Markov policy $\pi \colon \mathcal{S} \to \mathcal{A}$, which deterministically (i.e., not stochastically) selects an action only based on the state at the current time step. We denote an MDP under policy π as:

$$\mathcal{M}(\pi) \equiv \langle \mathcal{S}, \mathcal{A}, p_{s_0}, T, R, \pi \rangle \tag{2.8}$$

and the set of all the deterministic Markov policies as Π. Then the time evolution of an MDP $\mathcal{M}(\pi)$ can be obtained by the following steps:

Step 1. Let $t \leftarrow 0$ and initialize the initial state with the initial state probability function as $s_t \sim p_{s_0}$.

Step 2. Select action based on the current state as $a_t \leftarrow \pi(s_t)$.

Step 3. Take action a_t, receive reward $R(s_t, a_t)$, and transit to the next state s_{t+1} according to the probability distribution $T(s_t, a_t, s_{t+1})$.

Step 4. Let $t \leftarrow t + 1$ and go to **Step 2**.

The *expected reward* and the *expected discounted cumulative reward* are usually used as the objective function. Let us denote the reward the agent receives at time step t as a random variable R_t, then the expected reward is:

$$\text{The expected reward} \equiv \mathbb{E}\left[\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}R_t \;\middle|\; \mathcal{M}(\pi)\right] \tag{2.9}$$

and the expected discounted cumulative reward is:

$$\text{The expected discounted cumulative reward} \equiv \mathbb{E}\left[C_0 \mid \mathcal{M}(\pi)\right] \tag{2.10}$$

where C_t is the discounted cumulative reward:

$$C_t \equiv \lim_{K\to\infty}\sum_{k=0}^{K}\gamma^k R_{t+k} \tag{2.11}$$

where $\gamma \in [0,1)$ is called the discount rate and controls the virtual time horizon considered in the cumulative reward.

Let us consider a deterministic Markov policy π. For a given state s, value function $V^\pi\colon \mathcal{S}\to\mathbb{R}$ can be defined as follows:

$$\forall s\in\mathcal{S}\colon V^\pi(s) \equiv \mathbb{E}^\pi\left[C_0 \mid S_0 = s\right] \tag{2.12}$$

where $\mathbb{E}^\pi$ represents the expected value operator conditioned by the Markov chain defined by policy π, and S_0 is the initial state. If the resulted Markov chain has ergodic property (i.e., it is irreducible[2] and aperiodic[3]), the value function satisfies the following

[2]A Markov chain is irreducible if and only if $\forall s,s'\in\mathcal{S}\colon \exists t\in\mathbb{N}$ s.t. $\Pr(S_t=s'|S_0=s)>0$, i.e., any state is eventually reached from any other state.

[3]A Markov chain is aperiodic if and only if $\forall s\in\mathcal{S}\colon \gcd\mathcal{T}(s)=1$ where $\mathcal{T}(s)\equiv\{t\geq 1 \mid \Pr(S_t=s\mid S_0=s)>0\}$. For example, a Markov chain where a door has two states "open" and "closed" is *not* aperiodic because it is impossible to start from state "closed" and return to it with an odd number of transitions.

equation, known as the *Bellman equation*:

$$V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} T(s, \pi(s), s') V^{\pi}(s') \tag{2.13}$$

It is proven that there exists a deterministic Markov policy π that maximizes $V^{\pi}(s)$ in any state s and is called the optimal policy. Formally, we can define the optimal value function $V^{*}(s) = \max_{\pi \in \Pi}(V^{\pi}(s))$, then there exists a policy π^{*} that satisfies $\forall s \in \mathcal{S}\colon V^{\pi^{*}}(s) = V^{*}(s)$. The optimal value function satisfies the following recursive equation called *optimal Bellman equation*:

$$V^{*}(s) = \max_{a \in \mathcal{A}} \left[R(s, a) + \gamma \sum_{s' \in \mathcal{S}} T(s, a, s') V^{*}(s') \right] \tag{2.14}$$

Once the optimal Bellman equation is solved, the optimal policy can be defined as:

$$\pi^{*}(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \left[R(s, a) + \gamma \sum_{s' \in \mathcal{S}} T(s, a, s') V^{*}(s') \right] \tag{2.15}$$

One way to solve the optimal Bellman equation is Q-learning [69, 70], shown in Algorithm 1. In Q-learning, the agent learns the optimal action-value function $Q^{*}\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A}\colon Q^{*}(s, a) \equiv \max_{\pi \in \Pi} Q^{\pi}(s, a) \tag{2.16}$$

where Q^{π} is the action-value function:

$$Q^{\pi}(s, a) \equiv \mathbb{E}^{\pi}\left[C_0 \mid \; S_0 = s, A_0 = a \right] \tag{2.17}$$

Once the optimal action-value function is obtained, the optimal policy can also be obtained as:

$$\pi^{*}(s) = \operatorname*{argmax}_{a \in \mathcal{A}} Q^{*}(s, a) \tag{2.18}$$

Algorithm 1 Q-learning [69, 70]

Require: An environment with known $\mathcal{S}$ and $\mathcal{A}$, a policy model $\pi_t(a, s, q)$, a discount rate γ, a learning rate α_t, a termination condition, the number of episodes n
Ensure: Estimation of the optimal action-value function $\hat{Q}\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

Initialize $\hat{Q}(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, arbitrarily except that $\forall s \in \mathcal{S}_\mathrm{f}, a \in \mathcal{A}\colon \hat{Q}(s, a) = 0$ where $\mathcal{S}_\mathrm{f}$ is the set of all the terminal states.
for episode $= 1, n$ **do**
    Initialize the time step $t = 0$.
    Observe the initial state s_0 from the environment.
    repeat
        Select action a_t according to $\pi_t(a, s_t, \hat{Q})$ and take action a_t in the environment.
        Observe reward r_t and the next state s_{t+1} from the environment.
        Calculate the TD error: $\delta_t \leftarrow r_t + \gamma \max_{a' \in \mathcal{A}} \hat{Q}(s_{t+1}, a') - \hat{Q}(s_t, a_t)$
        Update $\hat{Q}$: $\hat{Q}(s_t, a_t) \leftarrow \hat{Q}(s_t, a_t) + \alpha_t \delta_t$
        Update the time step $t \leftarrow t + 1$
    until the termination condition is reached
end for

A policy model $\pi_t(a, s, q)$, one of the inputs to the Q-learning algorithm, defines which action the agent takes during the learning process based on the estimated optimal action-value function q at the time of the action. One of the common policy models is ε-greedy policy $\pi_\varepsilon\colon \mathcal{A} \times \mathcal{S} \times \mathbb{R}^{\mathcal{S}\times\mathcal{A}\times[0,1]} \to [0, 1]$. It is a stochastic Markov policy that selects a random action with probability ε, and the optimal action (at least under the assumption of $\hat{Q}$) with probability $1 - \varepsilon$. Formally,

$$\pi_\varepsilon(a, s, \hat{Q}, \varepsilon) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}|} & (\text{if } a = \operatorname{argmax}_{a'} \hat{Q}(s, a')) \\ \frac{\varepsilon}{|\mathcal{A}|} & (\text{otherwise}) \end{cases} \tag{2.19}$$

However, pure Q-learning is not applicable if the state space is continuous. Furthermore, even if the state space is discrete, it becomes computationally intractable if $|\mathcal{S}|$ is large. DeepMind [71] developed the Deep Q-Network (DQN) that approximates the optimal action-value function with a neural network with parameters θ. Two neural networks with the same architecture, the policy network Q_θ and the target network Q_{θ^-}, are prepared. The agent acts in the environment according to a policy model (e.g.,

the ε-greedy policy), and store each experience $e_t = (s_t, a_t, r_t, s_{t+1})$ in the experience replay memory $\mathcal{D}$. During the learning, the target network is updated to minimize the following loss at each step i:

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim U(\mathcal{D})}\left[\left(r + \gamma \max_{a'\in\mathcal{A}} Q_{\theta_i^-}(s', a') - Q_{\theta_i}(s, a)\right)^2\right] \tag{2.20}$$

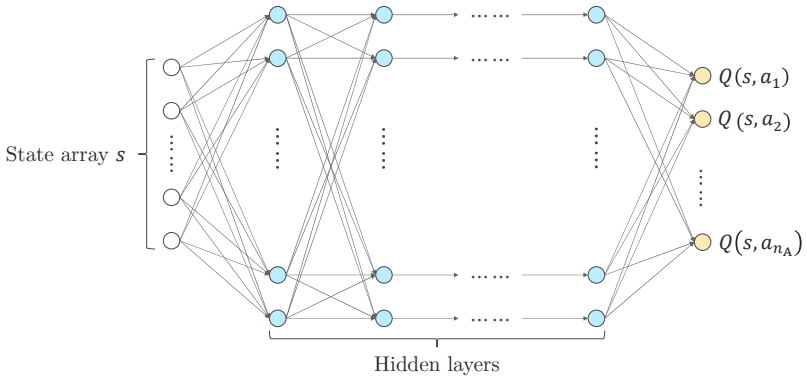
The expectation is calculated using samples drawn uniformly at random from the experience replay memory in the actual implementation. The target network parameters θ^- are updated with the policy network parameters θ every fixed number of steps. See Algorithm 2 for details. Wang et al. [72] developed the dueling network, where Q_θ is

Algorithm 2 Deep Q-Network [71]

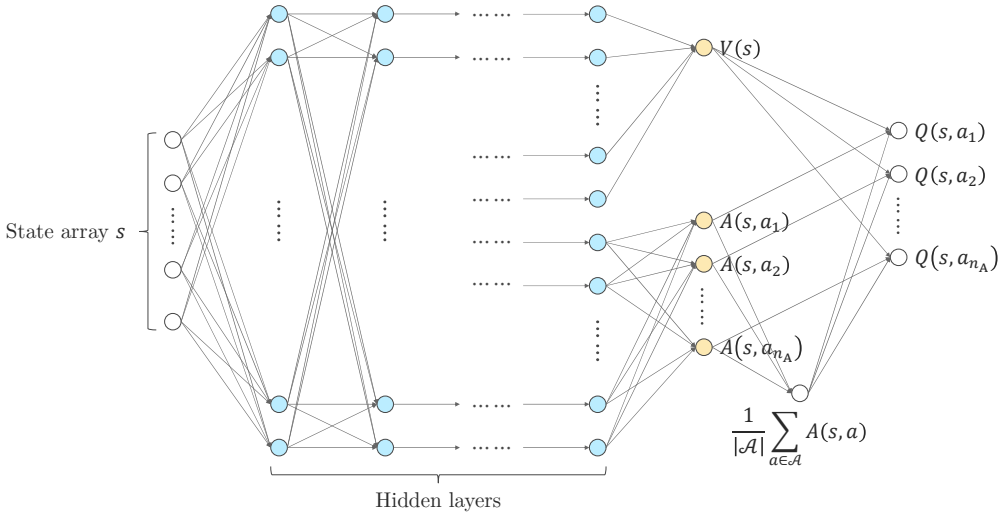
Require: An environment with known $\mathcal{S}$ and $\mathcal{A}$, a policy model $\pi(a, s, q)$, a discount rate γ, a network parameters optimizer, the replay memory capacity N, the target network update frequency C, a termination condition

Ensure: The approximated optimal action-value function Q_{θ^-}

Initialize the replay memory $\mathcal{D}$ with capacity N
Initialize the policy network parameters Q_θ with random weights θ
Initialize the policy network parameters Q_{θ^-} with the same weights $\theta^- = \theta$
for episode $= 1, n$ **do**
    Observe the initial state s_0 from the environment
    repeat
        Select action a_t according to $\pi_t(s_t, a_t, Q_{\theta^-})$ and take action a_t in the environment
        Observe reward r_t and the next state s_{t+1} from the environment
        Store experience $e_t = (s_t, a_t, r_t, s_{t+1})$ in $\mathcal{D}$
        Sample random batch of experiences (s_j, a_j, r_j, s_{j+1}) from $\mathcal{D}$
        Calculate $y_j = \begin{cases} r_j & \text{(if } s_{j+1} \text{ is a terminal state)} \\ r_j + \gamma \max_a Q_{\theta^-}(s_{j+1}, a) & \text{(otherwise)} \end{cases}$
        Perform an optimization step on $(y_j - Q_\theta(s_j, b_j, a_j, w_j))^2$
        Every C steps update target network $\theta^- = \theta$
    until the termination condition is reached
end for



(a) A Q-network



(b) A dueling Q-network

Figure 2.2 A single stream Q-network and the dueling Q-network.

decomposed into the value function $V(s)$ and the advantage function $A(s, a)$:

$$Q_\theta(s, a) = V_\theta(s) + \left(A_\theta(s, a) - \frac{1}{|\mathcal{A}|} \sum_{a' \in \mathcal{A}} A_\theta(s, a') \right) \tag{2.21}$$

The neural network architectures for the pure Deep Q-Network, and the dueling network are shown in Figure 2.2.

2.2 Decision-making under deep uncertainty

2.2.1 Robust decision making

Decision-makers often seek to predict the future and make a decision that performs the best in the predicted future. This approach is called "agree-on-assumptions." However, when making decisions in a fast-changing, complex world full of deep uncertainty, this approach is counter-productive and sometimes dangerous [44]. Robust decision making (RDM) [60] uses computer models not to predict the future but to simulate candidate policies in a wide range of plausible futures. Then the decision-maker can analyze the simulation results to observe the robustness and vulnerability of the policies. This approach is called "agree-on-decisions." RDM is "a set of concepts, processes, and enabling tools that use computation, not to make better predictions, but to yield better decisions under conditions of deep uncertainty [60]," and has four key elements [44]:

- Consider ensembles of a large number of scenarios.
- Seek robust, rather than optimal strategies.
- Employ adaptive strategies to achieve robustness.
- Use the computer to facilitate human deliberation over explorations, options, and trade-offs, not as a device for recommending a particular ordering of strategies.

As shown in Figure 2.3, an RDM analysis consists of the following steps.

Step 1. Decision structuring. The decision-makers define the key factors in the problem to analyze. This process often uses the "XLRM framework," where the decision-maker identify external factors (X), policy levers (L), relationship in the system (R), and performance metrics (M).

Step 2. Case generation. The decision-makers use simulation models to evaluate proposed strategies in a wide range of plausible futures.

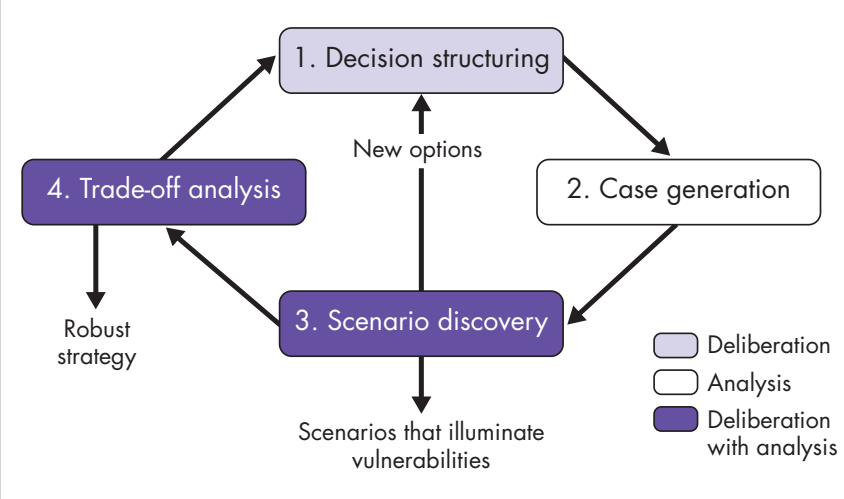


Figure 2.3 Iterative, participatory steps of an RDM analysis [60].

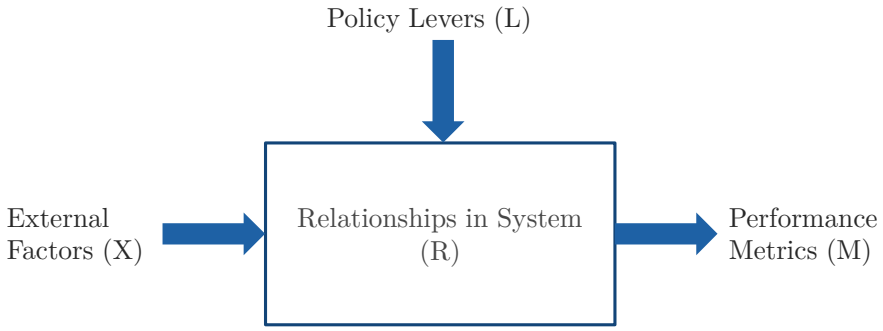


Figure 2.4 XLRM framework

Step 3. Scenario discovery. The decision-makers use visualization and data analysis methods to analyze the vulnerability of each policy. Scenario discovery is one of the commonly used analysis methods and identifies the key factors that affect on the performance metrics. The decision-makers may find new policy options from the results by, for example, synthesizing two policies with different advantages.

Step 4. Trade-off analysis. The decision-makers discuss which policy to adopt based on the scenario discovery and the trade-off analysis. If no policy satisfies the criteria, they can start over the process with new policy options or a renewed model.

Figure 2.5 is an example output of an RDM analysis [60]. In 2007, the U.S. Congress began to debate whether to reauthorize the Terrorism Risk Insurance Act (TRIA), passes in 2002 in the aftermath of the terrorist attacks in 2001. However, it was difficult

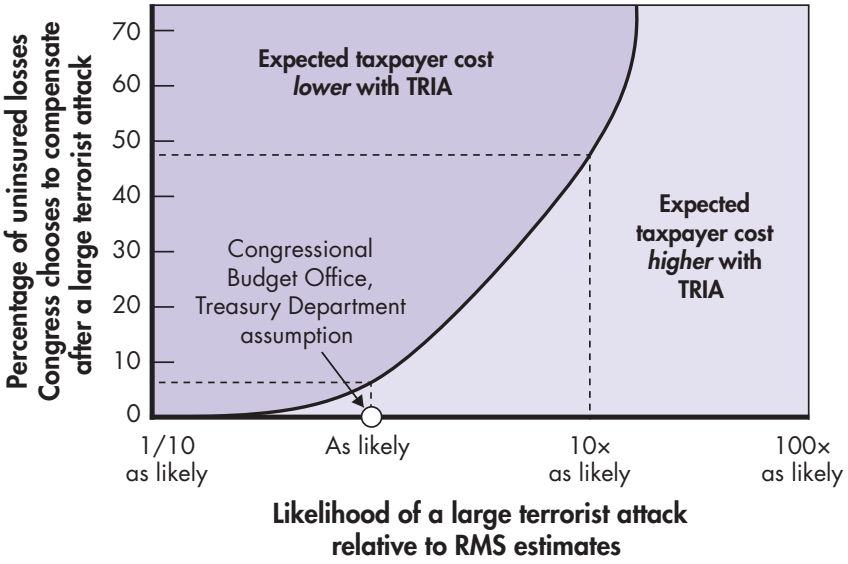


Figure 2.5 Expected annual taxpayer cost with and without TRIA [60].

to see whether the legislation would save taxpayers' money due to large uncertainty in various factors, including the likelihood of large terrorist attacks. RAND applied the RDM analysis by building a computer model and running simulations in various future scenarios. Among the 17 uncertain parameters, they found two key factors to determine whether the legislation would save taxpayers' money: the likelihood of a large terrorist attack and the amount that Congress compensates the uninsured. As shown as a white point in Figure 2.5, under the previous point assumption, the legislation was considered to be ineffective. However, the analysis showed that the legislation could be effective in a large area of scenarios, especially if the amount that Congress compensates the uninsured is large enough. This result made a significant contribution to the congressional debate.

2.2.2 Info-gap decision theory

The info-gap decision theory [61] is a design framework under deep uncertainty that aims to optimize the "robustness" of the decision rather than its performance. A problem is modeled with a vector of design variables $x \in \mathcal{X}$, a vector of uncertain

parameters $w \in \mathcal{W}$, and the objective function $f\colon \mathcal{X} \times \mathcal{W} \to \mathbb{R}$ to maximize. The decision-maker defines the info-gap model of uncertainty $\mathcal{U}(h)\colon \mathbb{R} \to 2^{\mathcal{W}}$. The info-gap model of uncertainty $\mathcal{U}(h) \subseteq \mathcal{W}$ defines a set of uncertain parameters that are considered to be possible based on the horizon of uncertainty h — a measure of to what extent uncertainty is considered. The uncertainty model $\mathcal{U}(h)$ generally has two properties: *contraction* and *nesting*. When $h = 0$, $\mathcal{U}(0)$ generally contains only one instance of uncertain parameters, the nominal value of the uncertain parameters denoted as $\tilde{w}$. This property is called *contraction*. If the horizon of uncertainty becomes greater, then the model contains more instances. Formally, if $h_1 \leq h_2$, then $\mathcal{U}(h_1) \subseteq \mathcal{U}(h_2)$. This property is called *nesting*.

The decision-maker then determines the critical value $r_{\mathrm{c}} \in \mathbb{R}$, the value that the objective function is required to exceed, such as the system performance requirement. The *info-gap robustness* $\hat{\alpha}(x, r_{\mathrm{c}})$ is the robustness measure in the info-gap decision theory. It is defined for a given design vector x and the critical value r_{c} as the maximum horizon of uncertainty h that the objective function $f(x, w)$ is better (greater in case of a maximization problem) than the critical value r_{c} for any uncertain parameter vector w in the info-gap model $\mathcal{U}(h)$. Formally,

$$\hat{\alpha}(x, r_{\mathrm{c}}) = \max \left\{ h \;\middle|\; \left[\min_{w \in \mathcal{U}(h)} f(x, w) \right] \geq r_{\mathrm{c}} \right\} \tag{2.22}$$

A large value of $\hat{\alpha}(x, r_{\mathrm{c}})$ indicates that under the design vector x, the critical value condition $f(x, w) \geq r_{\mathrm{c}}$ is satisfied even under uncertain parameter vectors considered to be far from the nominal value. Once the robustness is defined, then the design vector with the largest robustness can be defined as:

$$x^*_{\mathrm{ro}}(r_{\mathrm{c}}) = \operatorname*{argmax}_{x \in \mathcal{X}} \hat{\alpha}(x, r_{\mathrm{c}}) \tag{2.23}$$

Note that it is dependent on the critical value r_{c}.

In addition to the info-gap robustness, there is another measure called the *info-gap opportuneness*. The decision-maker determines the windfall value $r_{\mathrm{w}} \in \mathbb{R}$, the value that

the objective function could exceed, but only in limited cases. A windfall event could be an extra success of a space mission or a lottery win. The info-gap opportuneness is defined as the maximum horizon of uncertainty h that the objective function $f(x, w)$ is better than the windfall value for at least one uncertain parameter vector w in the info-gap model $\mathcal{U}(h)$. Formally,

$$\hat{\beta}(x, r_{\mathrm{w}}) = \max \left\{ h \;\middle|\; \left[\max_{w \in \mathcal{U}(h)} f(x, w) \right] \geq r_{\mathrm{w}} \right\} \tag{2.24}$$

A large value of $\hat{\beta}(x, r_{\mathrm{w}})$ indicates that under the design vector x, the windfall value condition $f(x, w) \geq r_{\mathrm{w}}$ is not satisfied until uncertain parameter vectors far from the nominal value are considered. Once the opportuneness is defined, then the design vector with the smallest opportuneness can be defined as:

$$x^*_{\mathrm{op}}(r_{\mathrm{w}}) = \operatorname*{argmin}_{x \in \mathcal{X}} \hat{\beta}(x, r_{\mathrm{w}}) \tag{2.25}$$

Note that it is dependent on the windfall value r_{w}.

2.3 Creativity support systems

Although being out of the scope of this research, creativity is essential for effective decision making. For example, it requires creative thinking to find a vague need, to embody it in concrete system requirements (i.e., a problem), and to propose candidates of designs (i.e., solutions) to be analyzed. In this sense, a creativity support system and a decision support system like the one this research proposes complement each other in the decision-making process. Figure 2.6 shows a typical decision-making process [73]. Creativity support systems can aid the initial steps: identifying the problem and generating alternatives, while the scope of this research is the next steps: evaluating and choosing alternatives. It has been reported that a creativity support system helps improve the process of, and outcome from, decision making [74, 75].

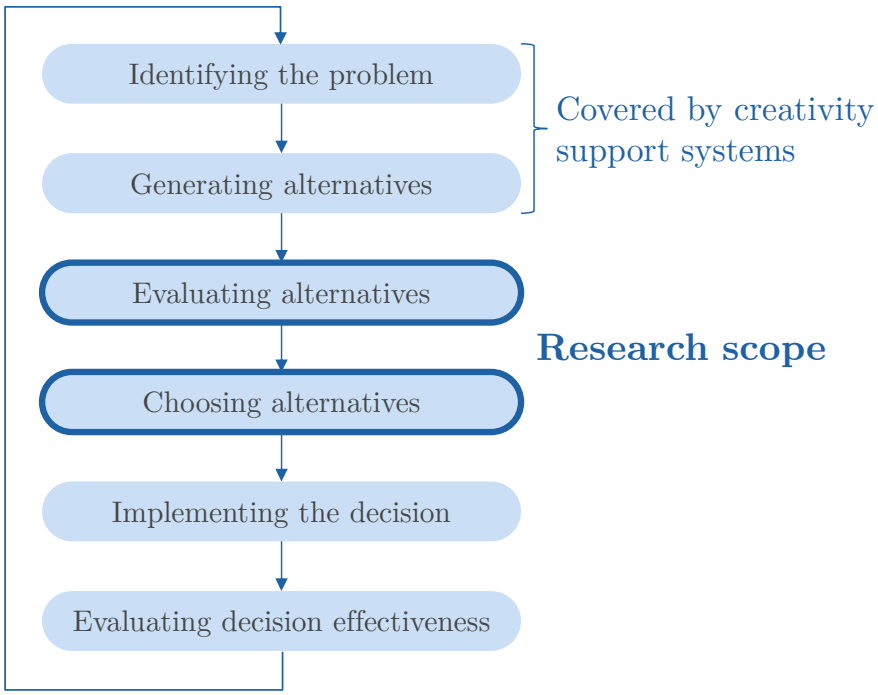


Figure 2.6 Typical decision-making process [73]. Creativity support systems can aid the first two steps: identifying the problem and generating alternatives. The scope of this research is the next two steps: evaluating alternatives and choosing alternatives.

As reviewed in [76], a creative process has stages, including problem finding, information finding, idea finding, and solution finding[4], and multiple creativity support systems have been proposed to support one or some of the stages. A notable example of a creativity support system is a creativity-enhancing decision-making support system (CDMSS) [74], shown in Figure 2.7. They conducted an experiment where participants were asked to make decisions on an airline's operation with a creativity enhancement tool called Axon Idea Processor, and the decision process and the outcome were compared with that of participants without it. They concluded that the CDMSS helped improve the process of, and outcome from, decision-making.

[4]Solution finding is different from idea finding in that complete solutions are produced by refining selected ideas and working out the details [76, 77].

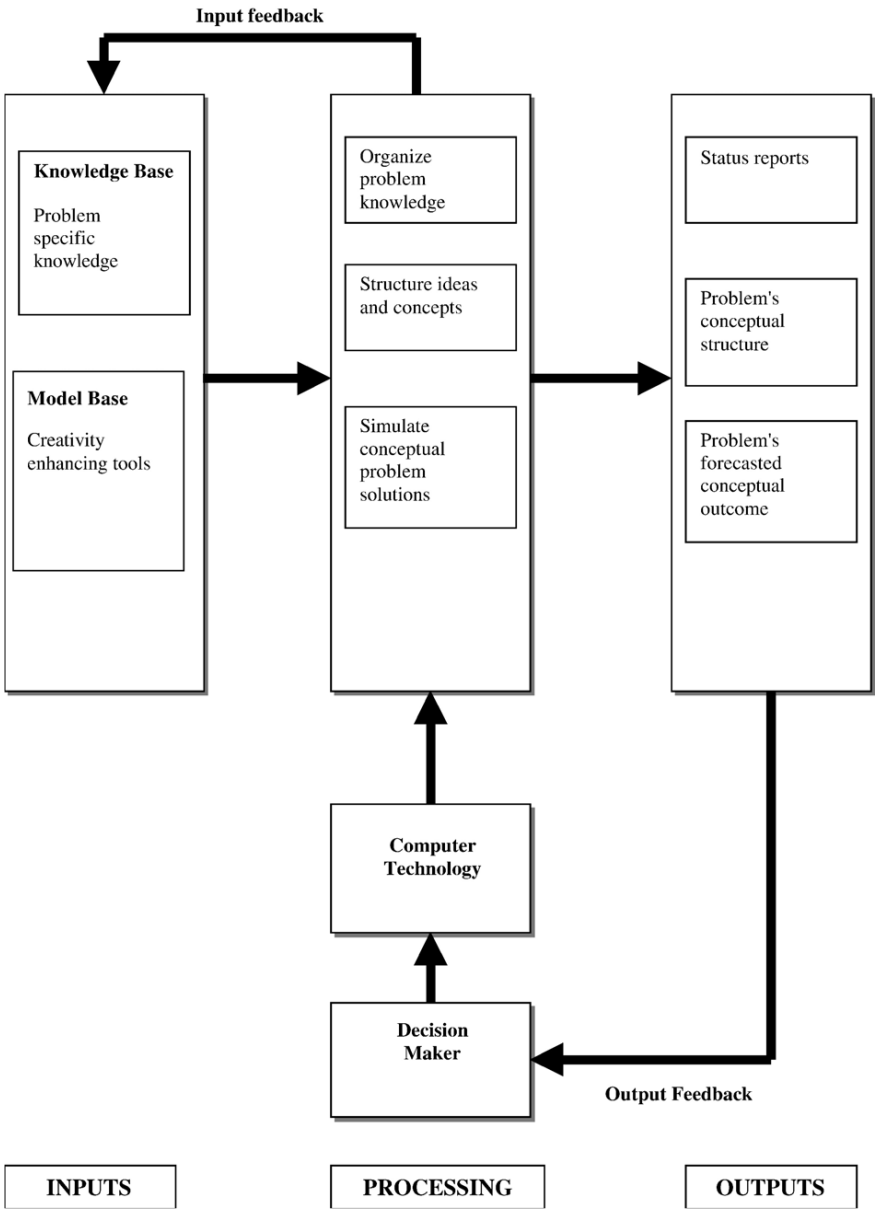


Figure 2.7 Creativity-enhancing decision-making support system (CDMSS) [74].

Chapter 3

Formulation of MSRDM-MDP

3.1 Definition of MSRDM-MDP

An MSRDM-MDP is an extension of an MDP in that it handles non-probabilistic uncertainties. In addition to the state set $\mathcal{S}$ and the action set $\mathcal{A}$, two sets are introduced: the scenario set $\mathcal{W}$ and the belief set $\mathcal{B}$. A scenario $w \in \mathcal{W}$ is a realization of the uncertain parameters. If there are d_{w} uncertain parameters $w_1, \ldots, w_{d_{\mathrm{w}}}$ where $w_i \in \mathcal{W}_i$, the scenario set is, or is a subset of, $\mathcal{W}_1 \times \cdots \times \mathcal{W}_{d_{\mathrm{w}}}$. The belief set $\mathcal{B}$ is defined as the set of all the subsets of $\mathcal{W}$. Formally, $\mathcal{B} \equiv 2^{\mathcal{W}}$. A belief $b \in \mathcal{B}$ is a subset of $\mathcal{W}$ and represents the set of scenarios that the agent considers possible.

An MSRDM-MDP $\mathcal{M}_{\mathrm{MSRDM}}$ can be defined as a tuple of six elements:

$$\mathcal{M}_{\mathrm{MSRDM}} \equiv \langle \mathcal{S}, \mathcal{W}, \mathcal{A}, R, T, \gamma, \rangle \tag{3.1}$$

where $\mathcal{S}$ is a set of all states, $\mathcal{W}$ is a set of all scenarios, $\mathcal{A}$ is a set of all actions, $R\colon \mathcal{S} \times \mathcal{A} \times \mathcal{W} \to \mathbb{R}$ is the reward function, $T\colon \mathcal{S} \times \mathcal{B} \times \mathcal{A} \times \mathcal{W} \to \mathcal{S} \times B$ is the transition function. $R(s, a, w)$ defines the reward the agent receives when it takes action a in state s under scenario w, and $(s', b') = T(s, b, a, w)$ defines the next state and belief after the agent transit from state s and belief b by taking action a under scenario w.

3.2 Bellman equation

3.2.1 Assumptions

Let us consider a transition from state s_t and belief b_t to state s_{t+1} and belief b_{t+1} by action a_t under scenario w, receiving reward r_t, i.e., $(s_{t+1}, b_{t+1}) = T(s_t, b_t, a_t, w)$ and $r_t = R(s_t, a_t, w)$. The transition function should be defined so that it satisfies the following properties:

$$w \in b_{t+1} \tag{3.2a}$$

$$b_{t+1} \subseteq b_t \tag{3.2b}$$

$$\forall w' \in b_{t+1} \colon T_{\mathrm{s}}(s_t, b_t, a_t, w') = s_{t+1} \wedge R(s_t, a_t, w') = r_t \tag{3.2c}$$

where $T_{\mathrm{s}} : \mathcal{S} \times \mathcal{B} \times \mathcal{A} \times \mathcal{W} \to \mathcal{S}$ is the state transition function, formally defined as:

$$T_{\mathrm{s}}(s, b, a, w) \equiv s' \text{ where } (s', b') = T(s, b, a, w) \tag{3.3}$$

Equation (3.2a) constrains the new belief to contain the true scenario, assuming the agent does not have false beliefs, Equation (3.2b) forbids a belief to expand, assuming an unexpected scenario outside the agent's belief does not happen, and Equation (3.2c) ensures that the state transition and reward are justified in all the scenarios in the new belief.

Another assumption is that the time horizon is finite. We denote the time horizon as T.

3.2.2 Policy and objective function

Once an MSRDM-MDP is defined, the next step is to find how the agent should act to receive as much reward as possible. To avoid loss of generality, we consider the set of all deterministic history-dependent policies. A history h_t consists of all the states,

beliefs, actions, and reward values from the start to time t and is formally defined as:

$$h_t \equiv (s_0, b_0, a_0, r_0, \ldots, s_{t-1}, b_{t-1}, a_{t-1}, r_{t-1}, s_t, b_t) \in \mathcal{H}_t \tag{3.4}$$

and a deterministic history-dependent policy at time t is defined as a function $\pi_t^{\mathrm{h}} : \mathcal{H}_t \to \mathcal{A}$.

Given an initial state s_0, an initial belief b_0, and a deterministic history-dependent policies $\boldsymbol{\pi}^{\mathrm{h}} = (\pi_0^{\mathrm{h}}, \ldots, \pi_{T-1}^{\mathrm{h}})$, the discounted cumulative reward in scenario w can be calculated as:

$$V^{\boldsymbol{\pi}^{\mathrm{h}}}(s_0, b_0, w) \equiv \sum_{t=0}^{T-1} \gamma^t R(s_t, \pi_t^{\mathrm{h}}(h_t), w) \tag{3.5}$$

The maximax optimal policy $\boldsymbol{\pi}^{\mathrm{h}*}_{+}$ and the maximin optimal policy $\boldsymbol{\pi}^{\mathrm{h}*}_{-}$ are the ones that maximizes the best-case and the worst-case respectively, formalized as:

$$\forall s_0, b_0 \colon \boldsymbol{\pi}^{\mathrm{h}*}_{+} = \operatorname*{argmax}_{\boldsymbol{\pi}^{\mathrm{h}} \in \Pi^{\mathrm{HD}}} \max_{w \in b_0} V^{\boldsymbol{\pi}^{\mathrm{h}}}(s_0, b_0, w) \tag{3.6}$$

$$\forall s_0, b_0 \colon \boldsymbol{\pi}^{\mathrm{h}*}_{-} = \operatorname*{argmax}_{\boldsymbol{\pi}^{\mathrm{h}} \in \Pi^{\mathrm{HD}}} \min_{w \in b_0} V^{\boldsymbol{\pi}^{\mathrm{h}}}(s_0, b_0, w) \tag{3.7}$$

where Π^{HD} is the set of all deterministic history-dependent policies.

3.2.3 Derivation of Bellman equation

Although scenario w is time-constant in MSRDM-MDP, let us virtually consider time-variant scenario w_t. This is equivalent to a two-player Markov game $\mathcal{M}_{\mathrm{G}}$ where the board is represented by (s_t, b_t), player one (the agent) and player two (the world) select moves $a_t \in \mathcal{A}$ and $w_t \in b_t$ respectively in order, and the agent receives the reward $R(s_t, a_t, w_t)$.

It can be proved that the maximax/maximin optimal policy for player 1 in game $\mathcal{M}_{\mathrm{G}}$ is also the maximax/maximin optimal policy in the original MSRDM-MDP.

$$
\begin{aligned}
&\min_{w\in b_0} V^{\boldsymbol{\pi}^{\mathrm{h}}}(s_0, b_0, w) \\
&= \min_{w\in b_0} \Big[R\left(s_0, \pi_0^{\mathrm{h}}(h_0), w\right) + \gamma R\left(s_1, \pi_1^{\mathrm{h}}(h_1), w\right) + \cdots \\
&\quad \cdots + \gamma^{T-1} R\left(s_{T-1}, \pi_{T-1}^{\mathrm{h}}(h_{T-1}), w\right) \Big] \\
&\geq \min_{w_0\in b_0} \Bigg[R\left(s_0, \pi_0^{\mathrm{h}}(h_0), w\right) + \gamma \min_{w_1\in b_1} \Big[R\left(s_1, \pi_1^{\mathrm{h}}(h_1), w_1\right) + \cdots \\
&\quad \cdots + \gamma^{T-1} \min_{w_{T-1}\in b_{T-1}} \Big[R\left(s_{T-1}, \pi_{T-1}^{\mathrm{h}}(h_{T-1}), w\right) \Big] \cdots \Big] \Bigg] \\
&= R\left(s_0, \pi_0^{\mathrm{h}}(h_0), w_0^*\right) + \gamma R\left(s_1, \pi_1^{\mathrm{h}}(h_1), w_1^*\right) + \cdots \\
&\quad \cdots + \gamma^{T-1} R\left(s_{T-1}, \pi_{T-1}^{\mathrm{h}}(h_{T-1}), w_{T-1}^*\right) \\
&= R\left(s_0, \pi_0^{\mathrm{h}}(h_0), w_{T-1}^*\right) + \gamma R\left(s_1, \pi_1^{\mathrm{h}}(h_1), w_{T-1}^*\right) + \cdots \\
&\quad \cdots + \gamma^{T-1} R\left(s_{T-1}, \pi_{T-1}^{\mathrm{h}}(h_{T-1}), w_{T-1}^*\right) \\
&= V^{\boldsymbol{\pi}^{\mathrm{h}}}(s_0, b_0, w_{T-1}^*)
\end{aligned}
\tag{3.8}
$$

consider scenario w to be time-variant

The reward is justified in all the scenarios in the new belief (Equation (3.2c)).

For a maximax/maximin MDP, it is proven that there exists a deterministic stationary policy $\pi : \mathcal{S} \times \mathcal{B} \to \mathcal{A}$ that maximizes the maximum/minimum discounted cumulative reward. Therefore, the maximum/maximin optimal Bellman equations are respectively defined as:

$$V_+^*(s, b) = \max_{a\in\mathcal{A}} \max_{w\in b} \Big[R(s, a, w) + \gamma V_+^*(T(s, b, a, w)) \Big] \tag{3.9}$$

$$V_-^*(s, b) = \max_{a\in\mathcal{A}} \min_{w\in b} \Big[R(s, a, w) + \gamma V_-^*(T(s, b, a, w)) \Big] \tag{3.10}$$

Once the optimal Bellman equations are solved, the optimal policies for the best-case and the worst-case can be obtained by:

$$\pi_+^*(s, b) = \operatorname*{argmax}_{a\in\mathcal{A}} \max_{w\in b} \Big[R(s, a, w) + \gamma V_+^*(T(s, b, a, w)) \Big] \tag{3.11}$$

$$\pi_-^*(s, b) = \operatorname*{argmax}_{a\in\mathcal{A}} \min_{w\in b} \Big[R(s, a, w) + \gamma V_-^*(T(s, b, a, w)) \Big] \tag{3.12}$$

3.3 Solving the Bellman equations

Equations (3.11) and (3.12) can be approximated:

$$V_{+}^{*}(s,b) = \max_{a\in\mathcal{A}} \max_{w\in W(b)} \Big[R(s,a,w_i) + \gamma V_{+}^{*}(T(s,b,a,w))\Big] \tag{3.13}$$

$$V_{-}^{*}(s,b) = \max_{a\in\mathcal{A}} \min_{w\in W(b)} \Big[R(s,a,w_i) + \gamma V_{-}^{*}(T(s,b,a,w))\Big] \tag{3.14}$$

where $W(b)$ is a set of scenarios sampled from belief b. Note that $W(b)$ does not need to be *randomly* sampled but should be sampled to maximize or minimize the operand $R(s,a,w_i) + \gamma V_{\pm}^{*}(T(s,b,a,w))$.

A reinforcement learning algorithm can solve Equations (3.13) and (3.14). Algorithm 3 is the Deep Q-Network algorithm modified to solve maximin MSRDM-MDPs. Once the optimal state-belief-action-scenario function $Q_{\pm}^{*}(s,b,a,w)$ is solved, the optimal policies can be obtained by:

$$\pi_{+}^{*}(s,b) = \operatorname*{argmax}_{a\in\mathcal{A}} \max_{w\in W(b)} Q_{+}^{*}(s,b,a,w) \tag{3.15}$$

$$\pi_{-}^{*}(s,b) = \operatorname*{argmax}_{a\in\mathcal{A}} \min_{w\in W(b)} Q_{-}^{*}(s,b,a,w) \tag{3.16}$$

The neural network architecture can be a pure Q-network shown in Figure 3.1 or a dueling network shown in Figure 3.2.

Algorithm 3 Deep Q-Network modified for maximin MSRDM-MDPs

Require: An environment with known $\mathcal{S}$, $\mathcal{W}$, $\mathcal{A}$, and $\mathcal{B}$, a discount rate γ, a network parameters optimizer, the replay memory capacity N, the target network update frequency C, a termination condition

Ensure: The approximated optimal action-scenario-value function Q_{θ^-}

Initialize replay memory $\mathcal{D}$ with capacity N
Initialize policy network Q_θ with random weights θ
Initialize target network Q_{θ^-} with the same weights $\theta^- = \theta$
for episode $= 1, n$ **do**
    Initialize state s_0 and belief b_0
    for $t = 1, T$ **do**
        With probability ε select a random action a_t otherwise select $a_t = \operatorname{argmax}_a \min_{w \in W(b)} Q_\theta(s_t, b_t, a, w)$
        Execute action a_t and observe reward R_t, new state s_{t+1}, and new belief b_{t+1}
        Store transition $(s_t, b_t, a_t, w, R_t, s_{t+1}, b_{t+1})$ in $\mathcal{D}$
        Sample random batch of transitions $(s_j, b_j, a_j, w_j, R_j, s_{j+1}, b_{j+1})$ from $\mathcal{D}$
        Set $y_j = \begin{cases} r_j & \text{if } s_{j+1} \text{ is terminal} \\ r_j + \gamma \max_a \min_{w \in W(b)} Q_{\theta^-}(s_{j+1}, b_{j+1}, a, w) & \text{otherwise} \end{cases}$
        Perform a gradient descent step on $(y_j - Q_\theta(s_j, b_j, a_j, w_j))^2$
        Every C steps update target network $\theta^- = \theta$
    end for
end for

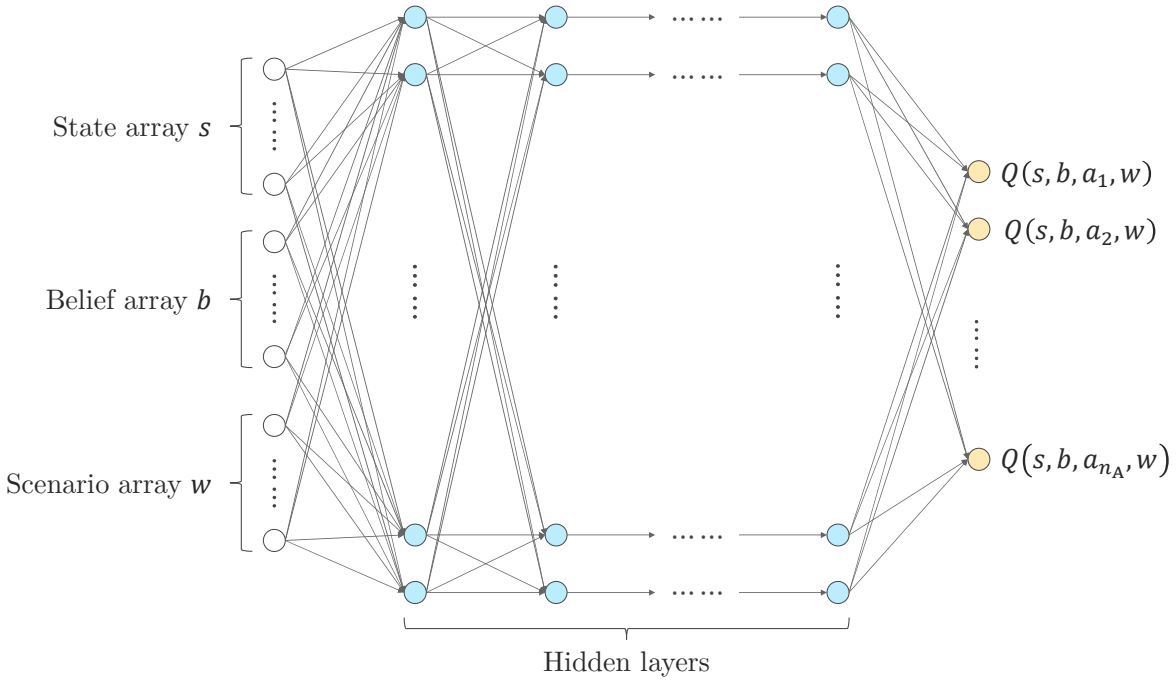


Figure 3.1 A Q-network for an MSRDM-MDP.

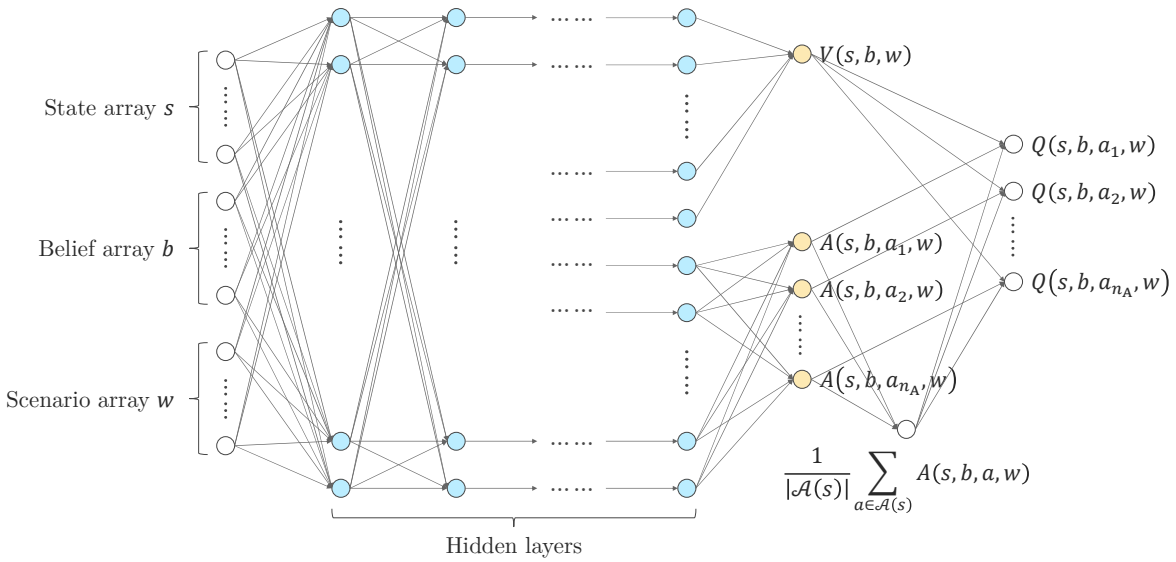


Figure 3.2 A dueling network for an MSRDM-MDP.

Chapter 4

Multi-Stage Robust Decision Making

4.1 MSRDM overview

Figure 4.1 shows an overview of the decision-making process using MSRDM, consisting of 5 steps:

1. **Decision structuring**

   In decision structuring, we formalize the problem in an MSRDM-MDP and express uncertainties as non-probabilistic uncertainty model $\mathcal{U}(h)$. To formulate the problem in an MSRDM-MDP, we need to identify:

   - Uncertain parameters (defined as "scenario" in the MSRDM-MDP). Examples include a technology's development costs and time, discrete policy direction scenarios, and the demand for a commodity.
   - Actions that the decision-maker can take (defined as "actions" in the MSRDM-MDP). Examples include to start developing a technology, to investigate an uncertain parameter's value, to deploy a technology, and to sell a product‘. Note that the definition of actions may limit the applicability

of reinforcement learning algorithms. For example, the deep Q-network (DQN) algorithm cannot be applied to problems with continuous actions.

- Variables that represent the state of the decision-maker at each time, are known to the decision-maker, and may change in time (defined as "state" in the MSRDM-MDP). Examples include a technology's time under development, whether a technology is completed or not, and a technology readiness level (TRL). Note that the change in state should be determined by the state itself, the action taken, and the true values of the uncertain parameters.
- How the "state" and the decision-maker's knowledge on the uncertain parameters ("belief") will change by each action under each realization of the uncertain parameters (defined as the "transition function" in the MSRDM-MDP).
- The performance measures (defined as the "reward function" in the MSRDM-MDP). Examples include technology development cost, realized technology performance, and the cumulative profit. Note that in the MSRDM-MDP, the "reward" should be a scalar value. Therefore, even when there are multiple performance measures, they should be represented by a scalar, for example, a weighted sum of each performance measure.

To construct the non-probabilistic uncertainty model, we need to identify:

- The nominal value of each uncertain parameter.
- The lower and upper bound of each uncertain parameter in the most extreme cases. Note that uncertainty is often assymetric. For example, a technology's development cost has a larger risk of exceeding the nominal (forecasted) value than that of being lower than the nominal value. It is also worth noting that uncertainty range estimated by experts often underestimate the uncertainty. Therefore, the lower and upper bounds should be defined conservatively not to exclude the unknown true value.

2. **Policy generation**

   In policy generation, we generate a set of policies by reinforcement learning and experts. Here, a "policy" is defined as a mapping from state–belief pair to an action. For example, a policy can be defined as "sell the product if the product is already developed and the worst-case demand is larger than some threshold, but not sell the product otherwise."

3. **Horizon of uncertainty (HoU) analysis**

   In HoU analysis, we calculate each policy's performance at different HoU values and show the robustness of the policy with the HoU plot. This step is conducted by a computer.

4. **Policy/HoU selection**

   We select a policy and HoU based on the HoU plot. The selected policy is simulated in the uncertainty set defined by the selected HoU.

5. **Scenario analysis**

   We simulate the selected policy under various realizations of the uncertain parameter vector in the defined uncertainty set and analyze the relationship between each uncertain parameter and the performance. We can apply various sensitivity analysis methods including feature scoring with machine learning regression algorithms (e.g., the extremely randomized tree), regional sensitivity analysis, and scenario discovery.

The following sections describe each step using the `SimpleMining` problem as an example.

4.2 Definition of the toy problem `SimpleMining`

Imagine two mines (mine 1 and mine 2) from which valuable resources can be extracted, and the agent can mine from only one of them. After mining from mine i, the agent receives a reward w_i whose value is uncertain. The uncertainty comes from each mine's

characteristics and the cost of preparing the necessary tools and mining operation. However, the agent can prospect a mine with a known prospecting cost to know the corresponding reward's exact value. The four actions that the agent can take are shown in Figure 4.2. The key questions the decision-makers have to answer are:

- Should they pay the prospecting cost to know the mining reward, or should they mine from one of the mines without prospecting?
- Which mine should they prospect or mine from?
- If they prospect a mine and know the reward, from which mine should they mine?

4.3 Decision structuring

4.3.1 Identifying relevant parameters using the XLRM framework

First, we need to identify external factors, policy levers, and performance metrics. External factors are exogenous parameters that affect the performance, and the decision-makers do not have complete control over them. Since the uncertainty in the problem is assumed to come from the uncertainty in the external factors, it is recommended that potential sources of uncertainty are defined as external factors. Note that the decision-makers may have *some* control over them. For example, future demand for a product is usually uncertain, but the company can control it to some extent by changing the effort and cost put into marketing.

In the `SimpleMining` problem, the external factors, policy levers, and performance metrics are defined as shown in Table 4.1. The external factors are the reward that the agent receives when mining from each mine. The policy levers are what action the agent takes in each situation. The agent can mine from one of the mines as well as prospect one. The performance metric is defined as the cumulative reward, which is

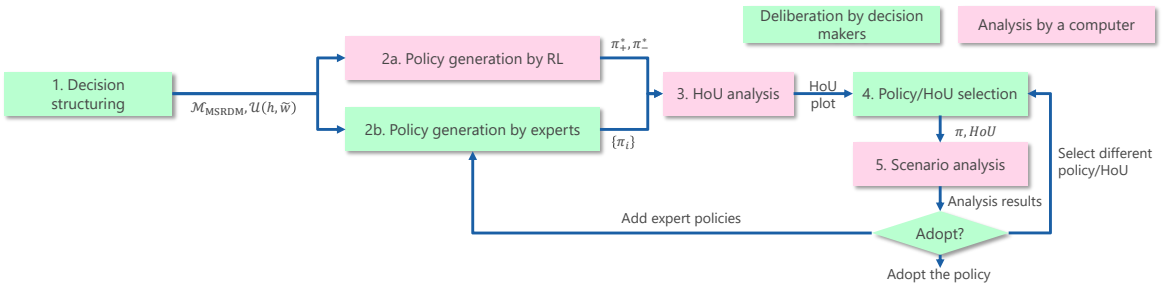


Figure 4.1 This is an overview of decision making using MSRDM. Green boxes are deliberation by decision-makers and red boxes are analyses by a computer.

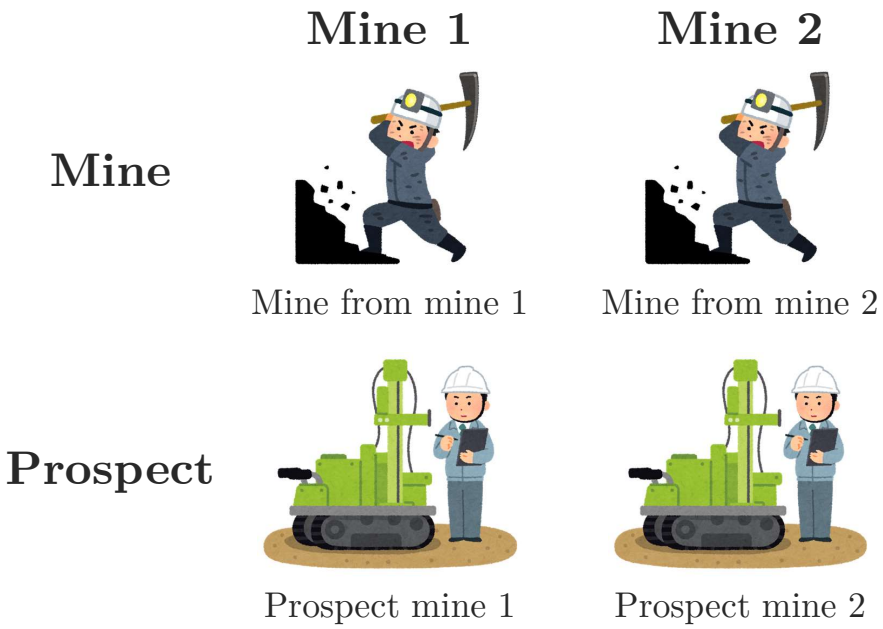


Figure 4.2 Four actions in the `SimpleMining` problem

Table 4.1 List of external factors, policy levers, and performance metrics in the `SimpleMining` problem.

External factors (X)	Reward from mine 1 Reward from mine 2
Policy levers (L)	Mine or prospect? Which mine to mine from or prospect?
Performance metrics (M)	Cumulative reward

identical to the mining reward in this problem because the agent can take a mining action only once. The relationship in the system (R) is later defined in Section 4.3.3.

4.3.2 Defining the non-probabilistic uncertainty model

To represent non-probabilistic uncertainty, we adopted the info-gap model of uncertainty [61]. With the horizon of uncertainty h, which defines the degree of uncertainty considered in the decision making, the set of scenarios considered in the decision making is defined by a non-probabilistic uncertainty model. A non-probabilistic uncertainty model $\mathcal{U}(h)$ maps the horizon of uncertainty and the nominal scenario to a set of scenarios and is formally defined as:

$$\mathcal{U}(h)\colon \mathbb{R}^{+} \times \mathcal{W} \to 2^{\mathcal{W}} \tag{4.1}$$

where h is the horizon of uncertainty.

The region that the uncertainty model defines can take several types of shapes. Van der Burg et al. adopted the ellipsoid-bound info-gap model [78] defined as:

$$\mathcal{U}(h) = \left\{ w \;\middle|\; [w - \tilde{w}]^{\mathsf{T}} V [w - \tilde{w}] \leq h^2 \right\} \tag{4.2}$$

where V is a positive definite real symmetric matrix that defines the ellipsoid's orientation and length along each axis. Intuitively, $[w - \tilde{w}]^{\mathsf{T}} V [w - \tilde{w}]$ is the "normalized" distance from the nominal scenario $\tilde{w}$ to the given scenario w. The model defines

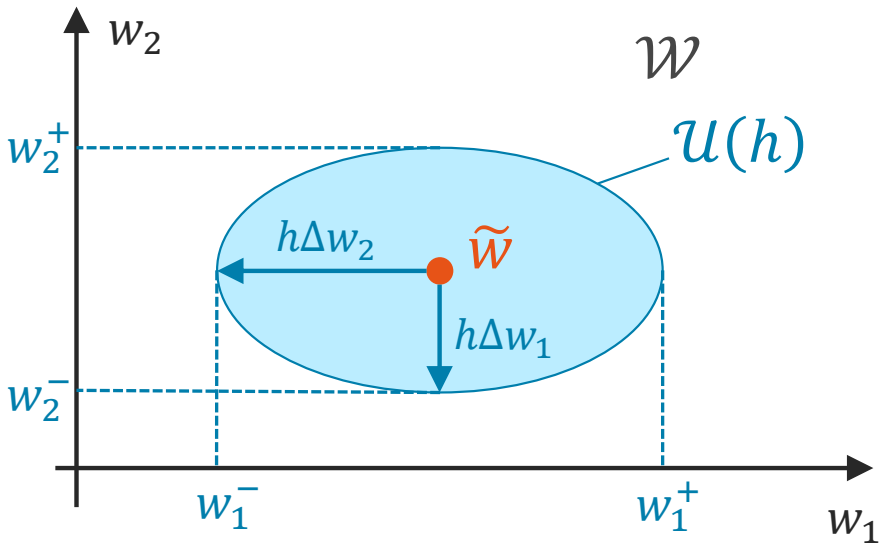


Figure 4.3 The elliptic uncertainty model $\mathcal{U}(h)$ for `SimpleMining`.

Table 4.2 Parameter values for the uncertainty model of `SimpleMining`.

Uncertain parameter	nominal value $\tilde{w}_i$	max deviation Δw_i
w_1, reward from mine 1	20	20
w_2, reward from mine 1	18	16

the uncertainty region as scenarios whose distance from the nominal is less than the threshold defined by the horizon of uncertainty h.

In the `SimpleMining` problem, the non-probabilistic uncertainty model for the uncertain parameters w_1, w_2 was defined as an elliptic uncertainty model shown in Figure 4.3. Formally,

$$\mathcal{U}(h) \equiv \left\{ (w_1, w_2) \;\middle|\; \left(\frac{w_1 - \tilde{w}_1}{\Delta w_1}\right)^2 + \left(\frac{w_2 - \tilde{w}_2}{\Delta w_2}\right)^2 \leq h^2 \right\}, \; 0 \leq h \leq 1 \tag{4.3}$$

where $\tilde{w}_i$ and Δw_i are the nominal value and the max deviation of the uncertain parameter w_i, respectively. Their values are shown in Table 4.2.

4.3.3 Defining the problem as an MSRDM-MDP

To define the problem as an MSRDM-MDP, the state set $\mathcal{S}$, the scenario set $\mathcal{W}$, the action set $\mathcal{A}$, the reward function R, the transition function T, and the discount factor γ are defined.

An MSRDM-MDP of `SimpleMining` can be formally defined as below:

- State set $\mathcal{S} = \{\texttt{NT}, \texttt{T}\}$ where $\texttt{NT}$ is the non-terminal state and $\texttt{T}$ is the terminal one.
- Scenario set $\mathcal{W} \subseteq \mathbb{R}^2$ where $w_i \in \mathbb{R}$ is the reward of mining from mine i.
- Action set $\mathcal{A} = \{\texttt{M}_1, \texttt{M}_2, \texttt{P}_1, \texttt{P}_2\}$ where $\texttt{M}_i$ is to mine from mine i, and $\texttt{P}_i$ is to prospect mine i.
- Reward function $R(s, a, w) = \begin{cases} 0 & (\text{if } s = \texttt{T}) \\ w_i & (\text{if } s = \texttt{NT} \wedge a = \texttt{M}_i) \\ -c_i & (\text{if } s = \texttt{NT} \wedge a = \texttt{P}_i) \end{cases}$
- Next state $s' = \begin{cases} \texttt{T} & (\text{if } s = \texttt{T} \vee a \in \{\texttt{M}_1, \texttt{M}_2\}) \\ \texttt{NT} & (\text{if } s = \texttt{NT} \wedge a \in \{\texttt{P}_1, \texttt{P}_2\}) \end{cases}$
- Next belief $b' = \{(w_1', w_2') \in b \mid \ w_i' = w_i\}$ where $i = \begin{cases} 1 & (\text{if } a = \texttt{M}_1, \texttt{P}_1) \\ 2 & (\text{if } a = \texttt{M}_2, \texttt{P}_2) \end{cases}$.

The definition of s' indicates that the agent transit to the terminal state after taking mining action and remains there ever after. The definition of b' indicates that when the agent mines from or prospect mine i, the value of w_i becomes known, and the belief is updated by discarding the scenarios in which w_i is not the same as the true value, as shown in Figure 4.4. One may find it incorrect that the lower and upper bounds of w_2 are changed by observing an independent random variable w_1. This paradox can be explained as an approximation to represent the value of prospecting actions. See Appendix A for the detailed discussion.

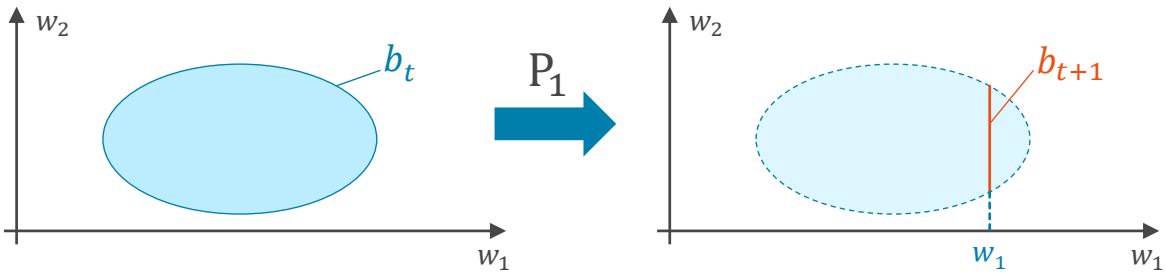


Figure 4.4 Update of belief b_t to b_{t+1} after taking action $\mathtt{P}_1$.

4.4 Policy generation

Once the MSRDM-MDP and the non-probabilistic uncertainty model are defined, the next step is to generate candidates of policies from which the decision-maker will select.

4.4.1 Policy generation by reinforcement learning

The maximax optimal policy π^*_+ and the maximin optimal policy π^*_- can be obtained using reinforcement learning, as described in Section 3.3.

4.4.2 Policy generation by experts

In addition to the maximax and maximin optimal policies generated by reinforcement learning, experts can also prepare explicitly expressed policies.

In the `SimpleMining` problem, the following three policies were defined:

Mine 1 The agent mines from mine 1.

Mine 2 The agent mines from mine 2.

Prospect The agent prospect mine 1. If the lower bound of w_2 in the new belief is larger than the true value of w_1, then the agent mines from mine 2, otherwise the agent mines from mine 1.

4.5 HoU analysis and policy/HoU selection

One of the analyses in the MSRDM framework is the horizon-of-uncertainty (HoU) analysis. For a horizon of uncertainty h, initial state s_0, initial belief b_0, and policy π, the maximum cumulative reward and the minimum cumulative reward in the set of scenarios $\mathcal{U}(h)$ can be calculated:

$$f^{+}_{\mathrm{HoU}}(h; s_0, b_0, \pi) = \max_{w \in \mathcal{U}(h)} V^{\pi}(s_0, b_0, w) \tag{4.4a}$$

$$f^{-}_{\mathrm{HoU}}(h; s_0, b_0, \pi) = \min_{w \in \mathcal{U}(h)} V^{\pi}(s_0, b_0, w) \tag{4.4b}$$

$V^{\pi}(s_0, b_0, w)$ is the cumulative reward the agent receives if it starts from state s_0 and belief b_0, and acts according to policy π under scenario w.

In the HoU plot, $f^{+}_{\mathrm{HoU}}(h; s_0, b_0, \pi)$ and $f^{-}_{\mathrm{HoU}}(h; s_0, b_0, \pi)$ is plotted for each policy π. Figure 4.5 shows the HoU plot for the `SimpleMining` problem. With the HoU plot, the decision-maker can visually understand which policy performs well even in the worst scenario and which policy performs well in the best scenario and consider the trade-off between performance and robustness.

Before moving to the next step of scenario analysis, the decision-maker can eliminate policies from the candidate policies if necessary. Also, the decision-maker needs to select the horizon of uncertainty to consider in the scenario analysis.

4.6 Scenario analysis

In the scenario analysis, scenarios are sampled from the uncertainty set $\mathcal{U}(h)$, and policies are simulated under each scenario. Let us denote the sampled scenarios as $\{w^{(j)}\}_{j=1}^{n_{\mathrm{w}}}$ where n_{w} is the number of scenarios.

There are several scenario analysis methods: the feature scoring, the scenario discovery, and the regional sensitivity analysis. In the feature scoring, the relationships between the performance metrics and the uncertain parameters in the scenario vector under each policy are regressed using a machine learning method such as extremely

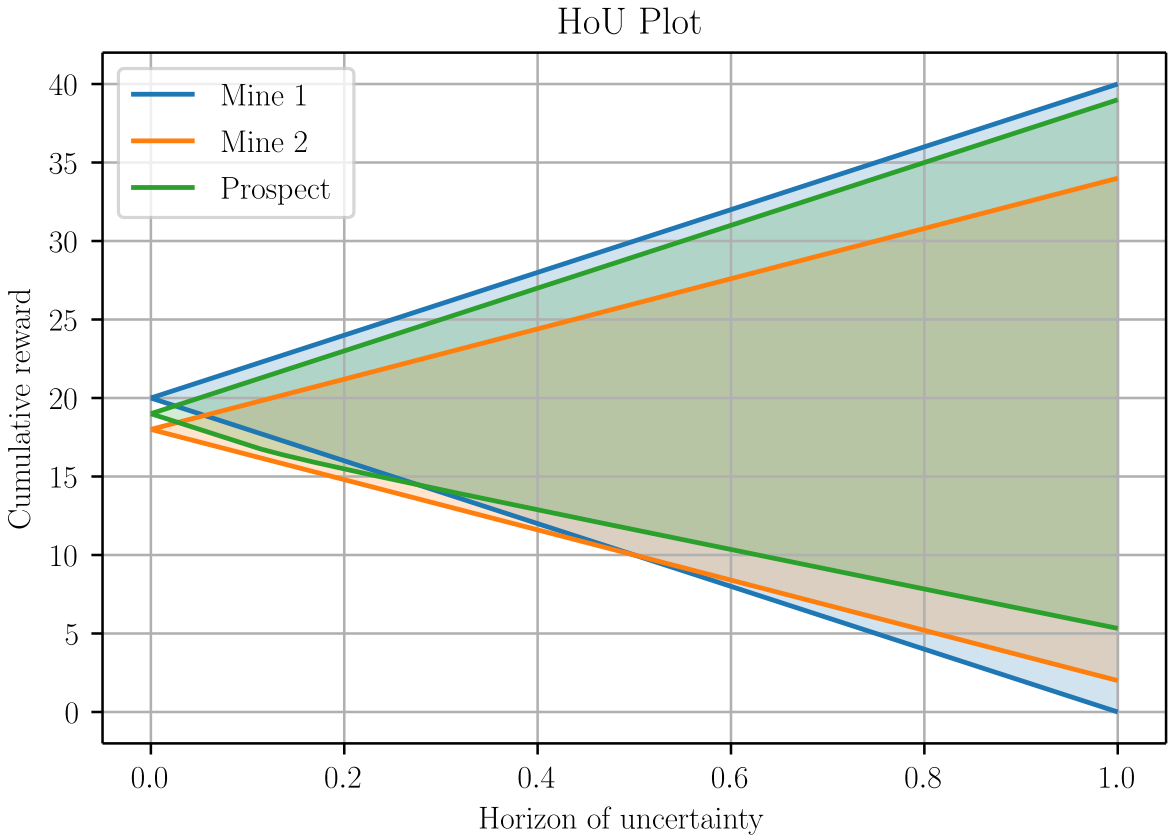


Figure 4.5 The HoU plot of the `SimpleMining` problem.

randomized trees [79]. In the scenario discovery, the decision-maker defines cases of interest (CoI). For example, one can define the CoI as cases where some performance target is achieved. Then boxes in $\mathcal{W}$ that contains the CoI can be obtained using the Patient Rule Induction Method (PRIM) [80].

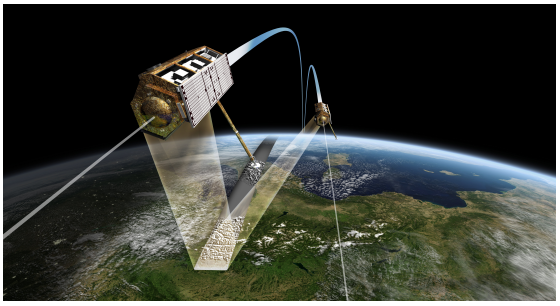
Chapter 5

Case Study I: Technology Roadmapping of Space Formation Flying System

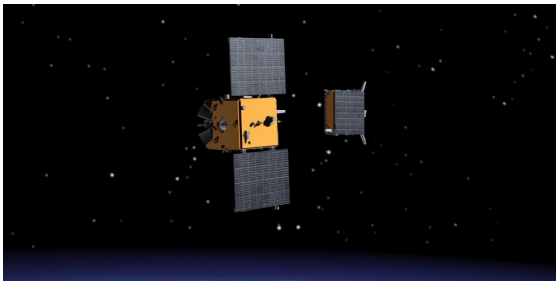
5.1 Background

The space formation flying (or formation flight) system comprises multiple spacecraft whose relative position and attitude are controlled to realize a function that a single spacecraft cannot have, such as space-based interferometry with long baselines.

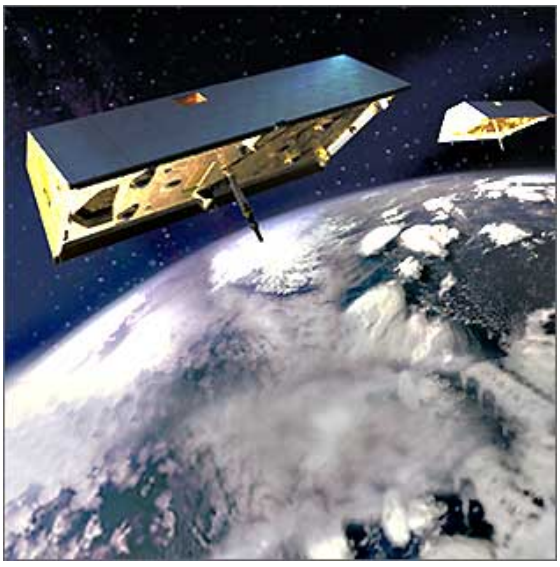
TanDEM-X [81] of DLR, launched in 2010, is a space-borne radar interferometer based on two TerraSAR-X radar satellites whose primary objective is generate a consistent global digital elevation model (DEM). The PRISMA [82], launched in 2010, was an experimental two-satellite mission to test formation-flying and rendezvous techniques. It demonstrated autonomous formation flying, and the control error in position was in the order of 0.1 to 1 m. The GRACE [83, 84] is a joint mission between NASA and DLR with two satellites launched in 2002, investigating Earth's gravity field. The two satellites fly at the altitude of 300 km to 500 km with a relative along-track separation of (220 ± 50) km.



(a) TanDEM-X (Credit: DLR [85])



(b) PRISMA [82]



(c) The GRACE [83]

Figure 5.1 Images of formation flying missions.

Table 5.1 Required functions for each mission concept. The checkmark (✓) represents whether the function is required for each mission: X-ray interferometry (`X`), infrared interferometry (`IR`), and gravitational wave telescope (`G`). Each function is categorized into three core technology groups (`CORE1`, `CORE2`, and `CORE3`) according to their required technology level.

Function	X	IR	G	Technology group
Autonomous control of relative position	✓	✓	✓	CORE1
Autonomous FDIR	✓	✓	✓	CORE1
Precise (mm) relative position control		✓	✓	CORE2
Linear formation flying with three spacecraft		✓		CORE2
Long-range (100s m) formation flying		✓		CORE2
Triangular formation flying with three spacecraft			✓	CORE3
Precise optical system control (sub-µm to nm)			✓	CORE3
Long-range (100 km) formation flying			✓	CORE3

However, relative position control of spacecraft with accuracy in the order of millimeters and relative position control of optical systems with accuracy in the order of sub-micrometers to nanometers need to be achieved to carry out missions such as X-ray interferometry ([86, 87]), infrared interferometry ([88, 89]), or gravitational wave telescope ([90, 91]) [92]. Therefore, we focused on three scientific mission concepts that benefit from formation flying technology: X-ray interferometry (`X`), infrared interferometry (`IR`), and gravitational wave telescope (`G`), and applied the proposed decision-support framework to its technology roadmapping problem. While some technologies are required for all the missions, others are required only for some missions. The functions required for each mission are listed in Table 5.1.

5.2 Decision structuring

5.2.1 Identifying relevant parameters using the XLRM framework

The external factors, policy levers, and performance metrics of the `FormationFlying` problem were defined as shown in Table 5.2.

Table 5.2 List of external factors, policy levers, and performance metrics in the `FormationFlying` problem.

External factors (X)	Development cost of each technology Development time of each technology Time limit for mission completion
Policy levers (L)	Which technology to develop
Performance metrics (M)	Cumulative reward

We identified as the external factors, i.e., the uncertainties, the development time and cost of the technologies, and the time limit. Potential sources of uncertainty in the technology development include limited estimation capability, unexpected effort due to technical issues during the development process, and schedule slip due to the annual budget limit. The technical uncertainty is often positively skewed (right-tailed), and its probability distribution is often modeled using the log-normal distribution [93], whose probability density function is shown in Figure 5.2, because the risk of higher cost or more extended schedule than initially planned is usually larger than the risk of lower cost or shorter schedule. We considered the time limit to be uncertain because the formation flying missions may be discontinued if running too long without completing the expected missions.

The policy options in the `FormationFlying` problem are defined by the staging of the development from the low-level to the high-level formation flying system. As shown in Table 5.1, more advanced technologies will be required as the scientific mission objective moves from the X-ray interferometry to the infrared interferometry and to the gravitational wave observation. The decision-makers can thus develop the formation flying system that has all the capabilities needed for the three missions at once, or they can develop the formation flying system for the X-ray interferometry first, then upgrade the system for the infrared interferometry, and finally upgrade the system for the gravitational wave observation. We did not define an information-obtaining action like the prospecting actions in the `SimpleMining` problem.

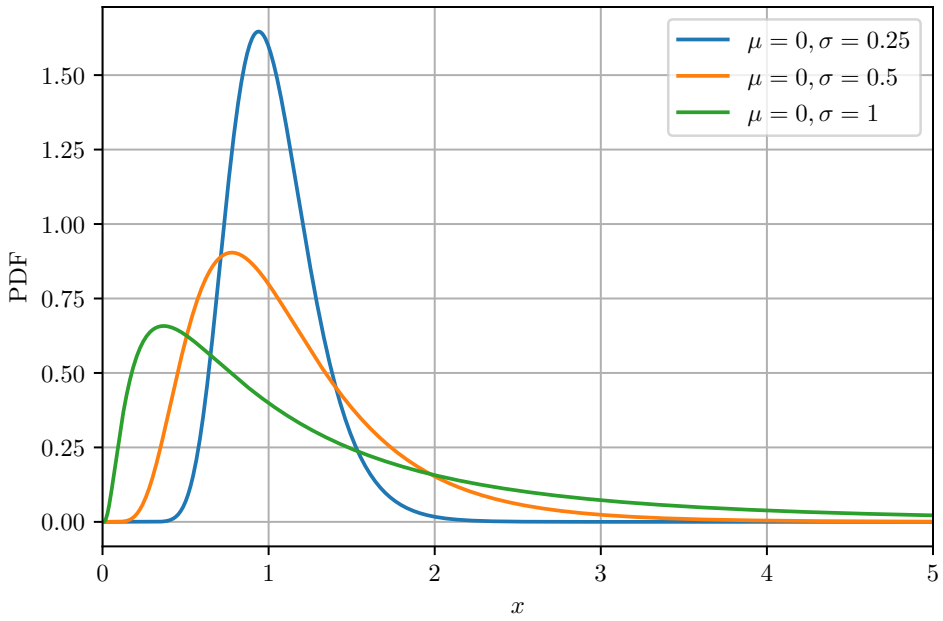


Figure 5.2 The probability density functions of the log-normal distributions with different μ and σ. If $x > 0$ is log-normally distributed, $\ln x$ is normally distributed with the mean μ and the standard deviation σ.

5.2.2 Defining the technologies and missions

In the `FormationFlying` problem, nine technologies were identified: `CORE1`, `CORE2`, `CORE3`, `CORE1to2`, `CORE1to3`, `CORE2to3`, `X`, `IR`, and `G`. `CORE1` is the formation flying technology required for the formation flying system designed to perform X-ray interferometry mission, which has two functions: autonomous control of relative position, and autonomous fault detection, isolation, and recovery (FDIR). `CORE2` is the formation flying technology required for the formation flying system designed to perform infrared interferometry mission, which has, in addition to the functions of `CORE1`, three functions: precise (in order of millimeters) relative position control, linear formation flying with three spacecraft, and long-range (in order of hundreds of meters) formation flying. `CORE3` is the formation flying technology required for the formation flying system designed to perform gravitational wave observation, which has, in addition to the functions of `CORE1` and `CORE2`, triangular formation with three spacecraft, precise optical system control (in order of sub-micrometer to nanometers), and long-range (in order of hundred kilometers) formation flying. `COREitoj` ($(i, j) \in \{(1, 2), (1, 3), (2, 3)\}$)

is the technological upgrade of the formation flying system from `COREi` to `COREj`, as visualized in Figure 5.3. `X`, `IR`, and `G` are the technology required to build the scientific instrument for the corresponding scientific missions.

Let n the number of technologies, and m the number of missions. Note that the technology names (`CORE1`, `CORE2`, ...) and their indices (1, 2, ...) are used interchangeably, and so are the mission names (`X`, `IR`, `X`) and their indices (1, 2, 3). We identified three missions and defined the mission feasibility function $f_{\text{feas},j}(\tau) : 2^{\{1,\ldots,n\}} \to \{\texttt{True}, \texttt{False}\} (j = 1, \ldots, m)$ representing whether the mission $\texttt{M}_j$ can be conducted with the set of technologies τ as:

$$f_{\text{feas},1}(\tau) = (\{\texttt{CORE1}, \texttt{X}\} \subseteq \tau) \vee (\{\texttt{CORE2}, \texttt{X}\} \subseteq \tau) \vee (\{\texttt{CORE3}, \texttt{X}\} \subseteq \tau) \tag{5.1a}$$

$$\begin{aligned} f_{\text{feas},2}(\tau) = {} & (\{\texttt{CORE2}, \texttt{IR}\} \subseteq \tau) \vee (\{\texttt{CORE1}, \texttt{CORE1to2}, \texttt{IR}\} \subseteq \tau) \\ & \vee (\{\texttt{CORE3}, \texttt{IR}\} \subseteq \tau) \vee (\{\texttt{CORE1}, \texttt{CORE1to3}, \texttt{IR}\} \subseteq \tau) \end{aligned} \tag{5.1b}$$

$$\begin{aligned} f_{\text{feas},3}(\tau) = {} & (\{\texttt{CORE3}, \texttt{G}\} \subseteq \tau) \vee (\{\texttt{CORE1}, \texttt{CORE1to2}, \texttt{CORE2to3}, \texttt{G}\} \subseteq \tau) \\ & \vee (\{\texttt{CORE1}, \texttt{CORE1to3}, \texttt{G}\} \subseteq \tau) \vee (\{\texttt{CORE2}, \texttt{CORE2to3}, \texttt{G}\} \subseteq \tau) \end{aligned} \tag{5.1c}$$

The definition of $f_{\text{feas},1}(\tau)$ indicates that mission 1 (`X`) can be conducted if `CORE1` and `X` are developed, `CORE2` and `X` are developed, or `CORE3` and `X` are developed.

5.2.3 Defining the non-probabilistic uncertainty model

The `FormationFlying` has 19 uncertain parameters: the time limit t_{lim}, by which all the missions should be completed, and the development cost c_i and development time T_i of each of the nine technologies. Here we denote a scenario as $w = (t_{\text{lim}}, c_1, \ldots, c_n, T_1, \ldots, T_n) \equiv (w_1, \ldots, w_{d_{\text{w}}})$, where $d_{\text{w}} = 2n+1 = 19$ is the number of uncertain parameters (i.e., the dimension of w). The uncertainty model $\mathcal{U}(h)$ $(0 \leq h \leq 1)$ was defined as an ellipsoid, except that it is asymmetric with respect to the nominal value. See Figure 5.5 for the visual image. The asymmetry captures the asymmetric uncertainty in technological development: the development cost tends to be higher than the original estimate rather than be lower, and so does the development time.

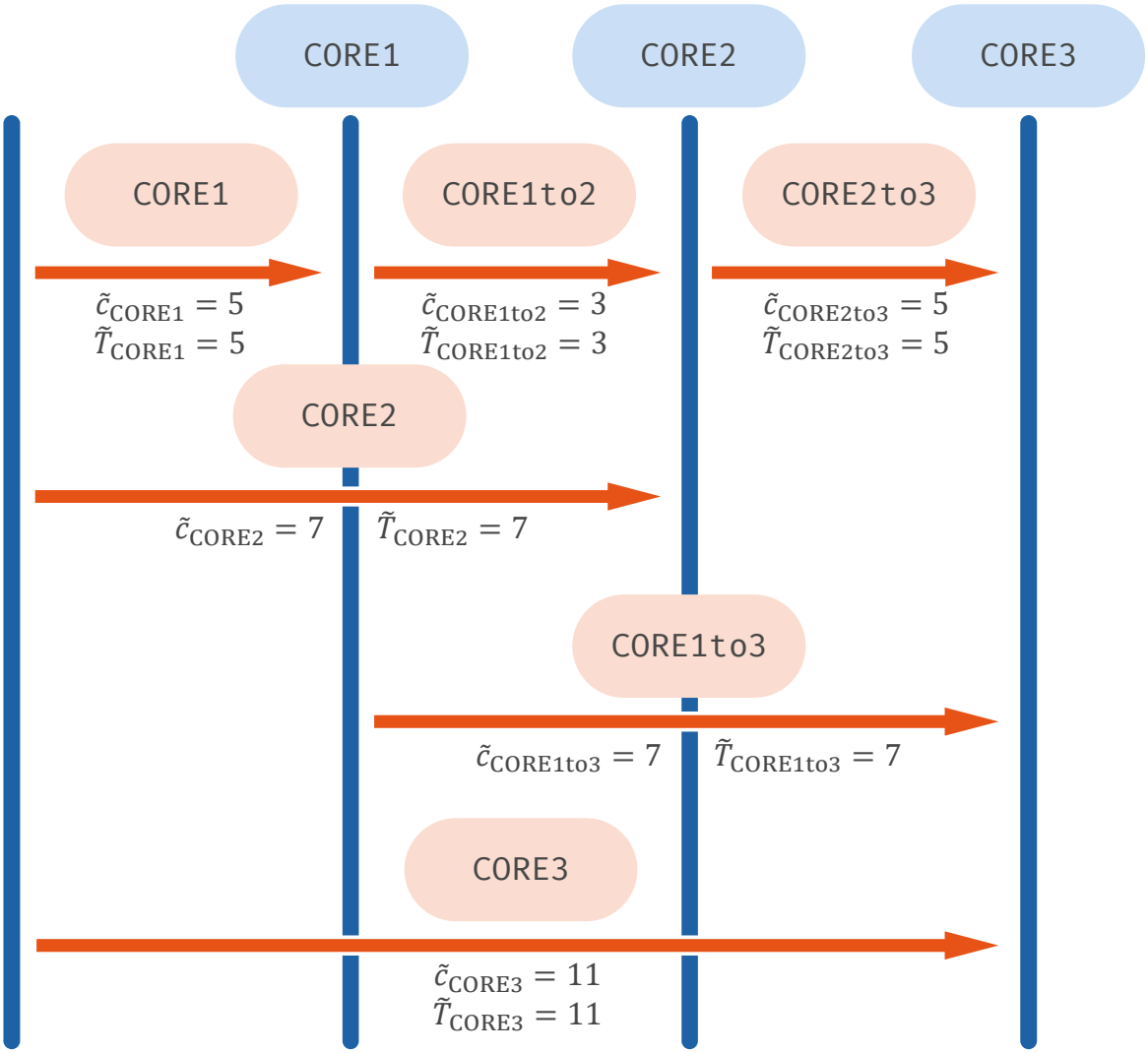


Figure 5.3 Definition of the formation flying core technologies. There are three levels of formation flying core technologies: `CORE1`, `CORE2`, and `CORE3`. Three upgrade technologies were defined in addition to the three core technologies.

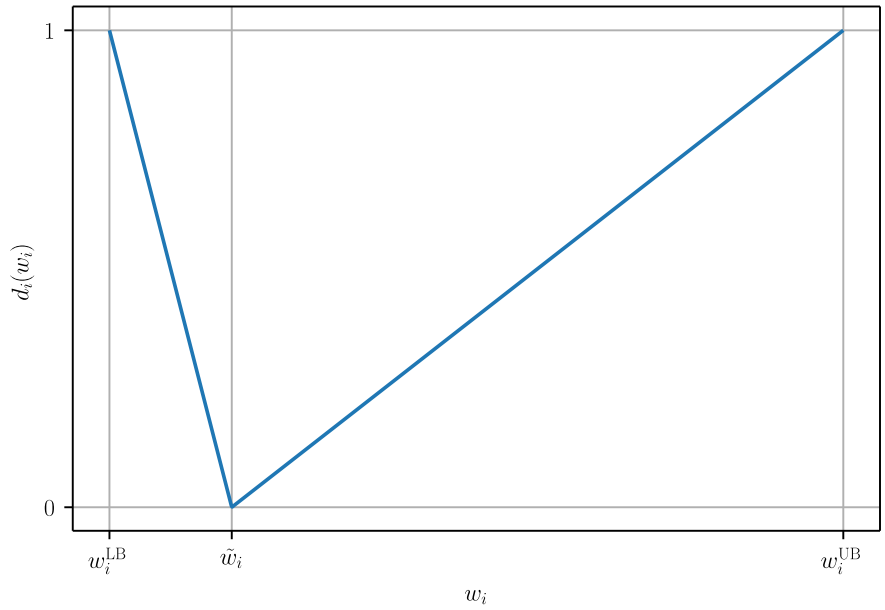


Figure 5.4 The distance function $d_i(w_i)$. The asymmetry of the function captures the uncertainty's skewness by considering a small deviation in the lower direction and a large deviation in the upper direction to be equally likely.

Formally, the uncertainty model was defined as:

$$\mathcal{U}(h) \equiv \left\{ (w_1, \ldots, w_{d_{\mathrm{w}}}) \;\middle|\; \sum_{i=1}^{d_{\mathrm{w}}} \left(d_i(w_i)\right)^2 \leq h^2 \right\}, \; 0 \leq h \leq 1 \tag{5.2}$$

where $d_i(w_i)$ is the distance function that represents the "distance" of the value w_i from the nominal value $\tilde{w}_i$:

$$d_i(w_i) = \begin{cases} \dfrac{\tilde{w}_i - w_i}{\tilde{w}_i - w_i^{\mathrm{LB}}} & (\text{if } w_i \leq \tilde{w}_i) \\ \dfrac{w_i - \tilde{w}_i}{w_i^{\mathrm{UB}} - \tilde{w}_i} & (\text{otherwise}) \end{cases} \tag{5.3}$$

where w_i^{LB} is the lower bound and w_i^{UB} is the upper bound of uncertain parameter w_i. See Figure 5.4 for the plot of the function. The asymmetry of the function captures the uncertainty's skewness by considering a small deviation in the lower direction and a large deviation in the upper direction to be equally likely.

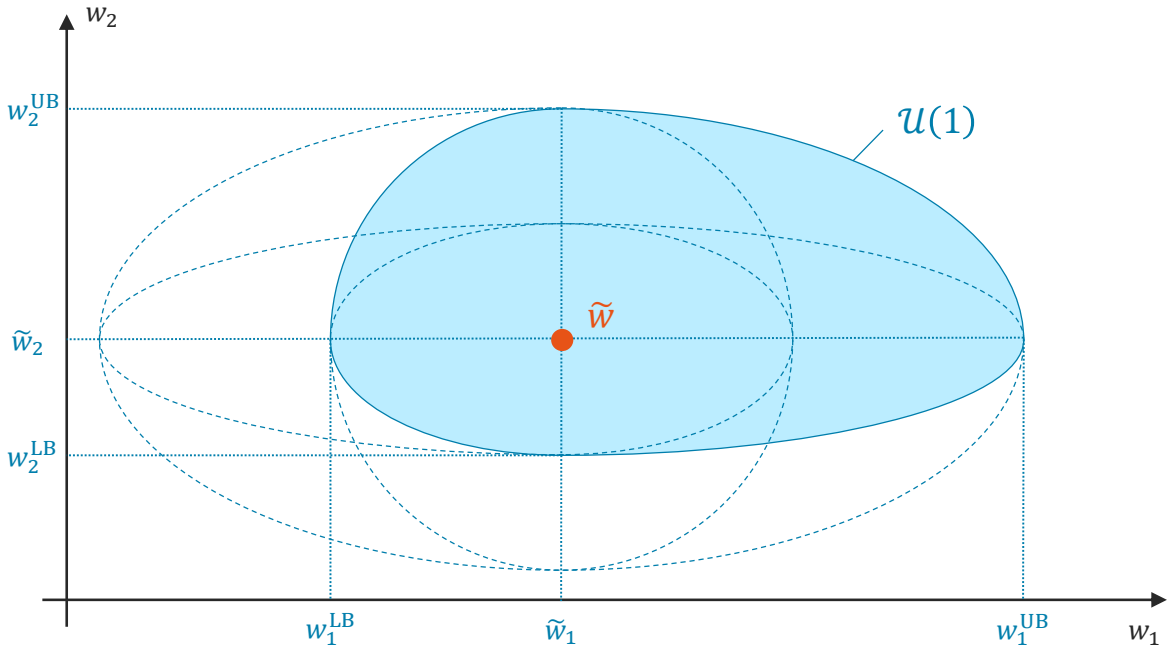


Figure 5.5 An example of an asymmetric uncertainty region $\mathcal{U}(1)$ when the dimension of w is two. The region can be divided into four subregions by the two lines $w_1 = \tilde{w}_1$ and $w_2 = \tilde{w}_2$, all of which are quartes of ellipses that share the same center point, the same orientation but have different principal axes, shown in the dashed curves.

The lower bound, the nominal value, and the upper bound of each uncertain parameter are defined in Table 5.3.

5.2.4 Defining the problem as an MSRDM-MDP

We defined an MSRDM-MDP of the `FormationFlying` problem as follows.

We defined a state as a vector $s = (t, u_1, \ldots, u_n, v_1, \ldots, v_n)$ where t is the current time, u_i is technology i's time under development, and v_i is a Boolean flag indicating whether the development of technology i is completed.

We defined a scenario as a vector $w = (t_{\mathrm{lim}}, c_1, \ldots, c_n, T_1, \ldots, T_n)$ where t_{lim} is the time limit, c_i is the development cost of technology i, and T_i is the development time of technology i.

We defined the action set as $\mathcal{A} = \{\mathtt{WAIT}, \mathtt{D}_1, \ldots, \mathtt{D}_n\}$ where `WAIT` is to do nothing in the current time step, and $\mathtt{D}_i$ is to start to develop technology i. Note that some

Table 5.3 Definitions of the uncertain parameters in the `FormationFlying` problem.

(a) The nominal value, the lower bound, and the upper bound of the development cost of each technology c_i. The percentage represents the deviation rate from the nominal value.

Technology	Nominal	Lower bound		Upper bound	
`CORE1`	5.0	4.0	(−20 %)	10.0	(+100 %)
`CORE2`	7.0	5.6	(−20 %)	14.0	(+100 %)
`CORE3`	11.0	8.8	(−20 %)	22.0	(+100 %)
`CORE1to2`	3.0	2.7	(−10 %)	4.5	(+50 %)
`CORE1to3`	7.0	6.3	(−10 %)	10.5	(+50 %)
`CORE2to3`	5.0	4.5	(−10 %)	7.5	(+50 %)
`X`	3.0	2.4	(−20 %)	6.0	(+100 %)
`IR`	3.0	2.4	(−20 %)	6.0	(+100 %)
`G`	3.0	2.4	(−20 %)	6.0	(+100 %)

(b) The nominal value, the lower bound, and the upper bound of the development time of each technology T_i. The percentage represents the deviation rate from the nominal value.

Technology	Nominal	Lower bound		Upper bound	
`CORE1`	5.0	4.0	(−20 %)	10.0	(+100 %)
`CORE2`	7.0	5.6	(−20 %)	14.0	(+100 %)
`CORE3`	11.0	8.8	(−20 %)	22.0	(+100 %)
`CORE1to2`	3.0	2.7	(−10 %)	4.5	(+50 %)
`CORE1to3`	7.0	6.3	(−10 %)	10.5	(+50 %)
`CORE2to3`	5.0	4.5	(−10 %)	7.5	(+50 %)
`X`	3.0	2.4	(−20 %)	6.0	(+100 %)
`IR`	3.0	2.4	(−20 %)	6.0	(+100 %)
`G`	3.0	2.4	(−20 %)	6.0	(+100 %)

(c) The nominal value, the lower bound, and the upper bound of time limit t_{lim}.

Nominal	Lower bound	Upper bound
20	10	30

actions may be invalid in some states. For state s, the set of valid actions $\mathcal{A}(s)$ was defined according to the following rules:

- `WAIT` is a valid action in all states.
- $\texttt{D}_i$ is a valid action in state s if the prerequisite technologies for technology i are already completed and the development of technology i has not been started.

The prerequisite technologies are defined only for the upgrade of formation flying technologies, namely `CORE1to2`, `CORE2to3`, and `CORE1to3`. These technologies cannot be developed unless the technologies from which each upgrade is made are completed. For instance, the agent cannot start developing `CORE2to3` unless a) `CORE2` is completed, or b) both `CORE1` and `CORE1to2` are completed. We defined the development readiness function $f_{\text{preq},i}(\tau)\colon 2^{\{1,\dots,n\}} \to \{\texttt{True}, \texttt{False}\}$ $(i = 1, \dots, n)$ that represents whether the prerequisite technologies for technology i are completed given a set of completed technologies τ. The set of valid actions $\mathcal{A}(s)$ can be written as:

$$\mathcal{A}(s) \equiv \{\texttt{WAIT}\} \cup \{\texttt{D}_i \mid \ i = 1, \dots, n;\ f_{\text{preq},i}(\tau(v))\} \tag{5.4}$$

where $\tau(v) \equiv \{i \mid \ v_i = \texttt{True}\}$ is the set of completed technologies.

We defined the reward function as

$$R(s, a, w) = -\sum_{i=1}^{n} \frac{c_i}{T_i} \left[a = \texttt{D}_i \vee (0 < u_i \wedge \neg v_i)\right] + \sum_{j=1}^{m} p_j \left[\neg f_{\text{feas},j}(\tau(v)) \wedge f_{\text{feas},j}(\tau(v'))\right] \tag{5.5}$$

where p_j is the reward for mission j defined in Table 5.4. Note that $[\cdot]$ in Equation (5.5) is the Iverson bracket[1]. The reward at each time step is the reward for the missions that become feasible at the time step minus the development cost of the technologies under development at the time step.

[1]The Iverson bracket $[\cdot]$ is a function that returns 1 if the statement within the brackets is true and 0 otherwise. Formally, $[P] = \begin{cases} 1 & \text{(if } P \text{ is true)} \\ 0 & \text{(otherwise)} \end{cases}$

Table 5.4 Reward for each mission

Index j	Mission	Reward (p_j)
1	X-ray interferometry (`X`)	40
2	Infrared interferometry (`IR`)	40
3	Gravitational wave telescope (`G`)	40

We defined the transition function in two steps: the state transition and the belief transition. The state transition defines the next state $s' = (t', u'_1, \ldots, u'_n, v'_1, \ldots, v'_n)$ as

$$t' = t + 1 \tag{5.6a}$$

$$u'_i = \begin{cases} \max\{u_i + 1, T_i\} & (\text{if } a = \mathtt{D}_i \vee u_i > 0) \\ u_i & (\text{otherwise}) \end{cases} \tag{5.6b}$$

$$v'_i = \begin{cases} 1 & (\text{if } u'_i \geq T_i) \\ 0 & (\text{otherwise}) \end{cases} \tag{5.6c}$$

If the time step reaches the time limit defined by the scenario, i.e., $t' \geq t_{\text{lim}}$, the next state s' is set to a terminal state, and the episode is terminated. If a technology development is completed, the agent will know the technology's development cost and time. This is modeled by the belief transition that defines the next belief b' as:

$$b' = \{(t'_{\text{lim}}, c'_1, \ldots, c'_n, T'_1, \ldots, T'_n) \in b \mid \ \forall i \colon (\neg v_i \wedge v'_i) \rightarrow c'_i = c_i \wedge T'_i = T_i\} \tag{5.7}$$

The state is initialized as $t = 0, u_i = 0, v_i = 0$ $(i = 1, \ldots, n)$. The belief is initialized with the horizon of uncertainty h as $b_0 = \mathcal{U}(h)$.

5.3 Policy generation

5.3.1 Policy generation by experts

We defined four expert policies:

Aggressive The agent starts to develop `CORE3`, `X`, `IR`, and `G`.

Staged (X, IR, G) The agent starts to develop `CORE1` and `X`. When `CORE1` is completed, it starts to develop `CORE1to2` and `IR`. When `CORE1to2` is completed, it starts to develop `CORE2to3` and `G`.

Staged (X/IR, G) The agent starts to develop `CORE2`, `X`, and `IR`. When `CORE2` is completed, it starts to develop `CORE2to3` and `G`.

Staged (X, IR/G) The agent starts to develop `CORE1` and `X`. When `CORE1` is completed, it starts to develop `CORE1to3`, `IR`, and `G`.

Note that in the MSRDM-MDP environment, the agent cannot take more than one action at one time step. Therefore, under the aggressive policy, the agent starts to develop `CORE3` at $t = 0$, `X` at $t = 1$, `IR` at $t = 2$, and `G` at $t = 3$. We visualized these policies in the technology roadmap format in Figure 5.6, showing the development timeline of each technology and the timeline of when each mission becomes feasible if all the uncertain parameters have their nominal values.

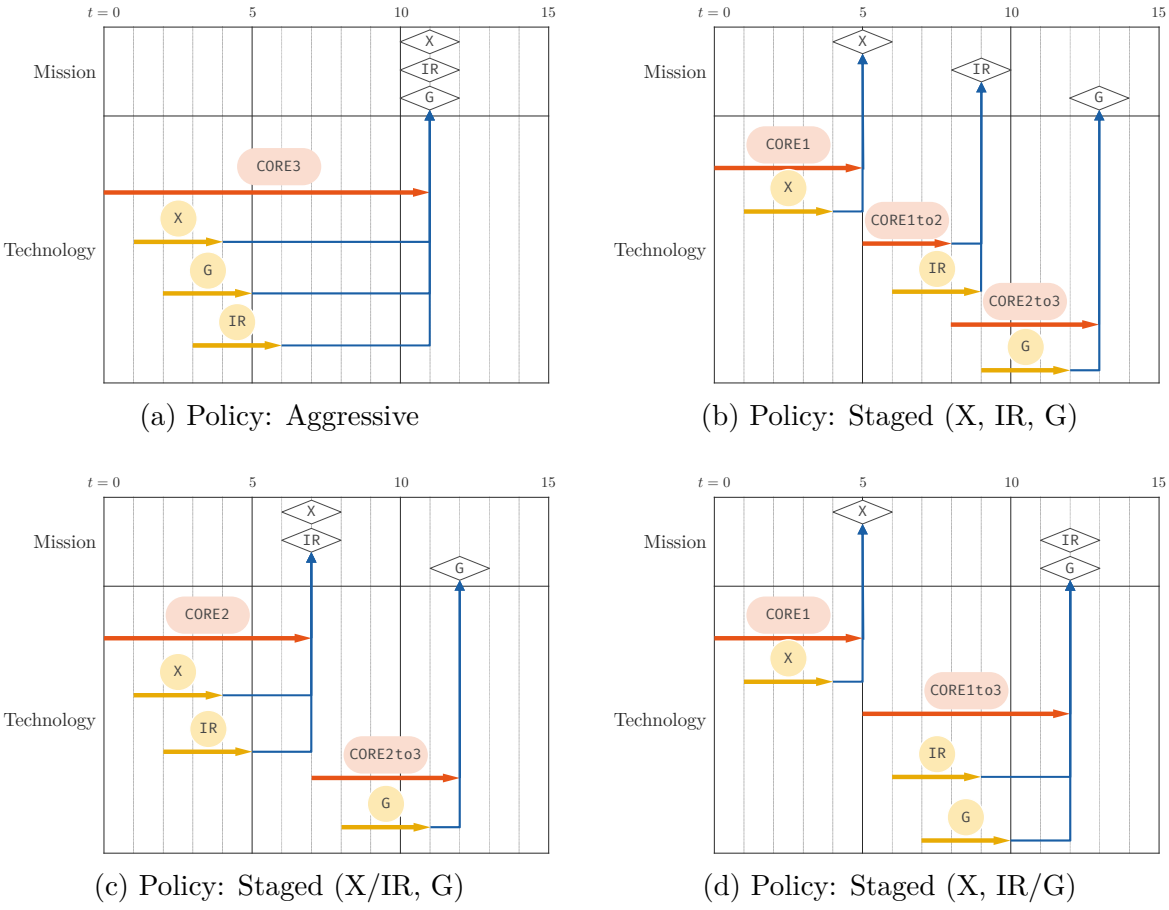
5.3.2 Policy generation by reinforcement learning

We did not apply reinforcement learning to the policy generation in the `FormationFlying` problem because all the possible staged development policies that enable all three missions are covered by the four expert policies defined in Section 5.3.1.

5.4 HoU analysis and policy/HoU selection

5.4.1 HoU analysis settings

The parameters used in the HoU analysis are shown in Table 5.5. 25 values of the horizon of uncertainty h were sampled with even spaces in $[0, 1]$ ($h = 0, \frac{1}{24}, \ldots, \frac{23}{24}, 1$). Under each horizon of uncertainty h, we sampled 1,000 scenarios from $\mathcal{U}(h)$, 19 of which were the Pareto vertices of the ellipsoid-like region $\mathcal{U}(h)$ and the others were



(a) Policy: Aggressive

(b) Policy: Staged (X, IR, G)

(c) Policy: Staged (X/IR, G)

(d) Policy: Staged (X, IR/G)

Figure 5.6 Technology roadmap for each expert policy. The nominal value is used for the development time of each technology.

Table 5.5 Parameters used in the HoU analysis of `FormationFlying` problem.

Parameter	Value
Discount rate γ	1
Scenarios sampling method	Uniform Pareto scenarios sampling and vertices sampling
The number of Pareto scenarios samples	981
The number of Pareto vertices samples	19 ($= d_{\mathrm{w}}$)
The samples of h	25 evenly spaced samples in $[0, 1]$

uniformly sampled from the Pareto-front (uniform Pareto scenarios sampling). In total, 25,000 scenarios were prepared.

Pareto vertices $\left\{w_{\mathrm{v}}^{(i)}(h)\right\}$ are scenarios that are located at vertices of the ellipsoid-like region $\mathcal{U}(h)$ and on the Pareto front, as illustrated in Figure 5.7. Formally, a Pareto vertex $w_{\mathrm{v}}^{(i)}(h)$ in calculating the minimum reward is:

$$w_{\mathrm{v}}^{(i)}(h) \equiv (\tilde{w}_1, \ldots, \tilde{w}_{i-1}, w_i^*(h), \tilde{w}_{i+1}, \ldots, \tilde{w}_{d_{\mathrm{w}}}) \tag{5.8}$$

where

$$w_i^*(h) \equiv \begin{cases} \tilde{w}_i - (\tilde{w}_i - w_i^{\mathrm{LB}})h & \text{(if } w_i\text{'s direction of goodness is positive.)} \\ \tilde{w}_i + (w_i^{\mathrm{UB}} - \tilde{w}_i)h & \text{(otherwise)} \end{cases} \tag{5.9}$$

The *direction of goodness* defines whether the uncertain parameter is preferred to be larger or smaller. In the `FormationFlying` problem, t_{lim}'s direction of goodness is positive, whereas the other uncertain parameters $(c_1, \ldots, c_n, T_1, \ldots, T_n)$ have the direction of goodness in the negative direction.

The Pareto scenarios are scenarios uniformly sampled from the Pareto front. Because the direction of goodness of each uncertain parameter is known, we can define the Pareto front in case of the cumulative reward minimization (maximization) as a set of all the scenarios where any of the uncertain parameters cannot be worse (better) without making any other parameter better (worse). In the `FormationFlying` problem, the Pareto front is the portion of the surface of a d_{w}-dimensional ellipsoid in the closed orthant containing all the Pareto vertices. The Pareto scenarios were uniformly sampled from the surface using a method developed by Marsaglia [94].

In addition to the four expert policies, we added a policy under which the agent always takes `WAIT` action. We simulated the five policies under the 25,000 scenarios, both for the maximum and minimum cumulative reward, resulting in 250,000 simulations in total.

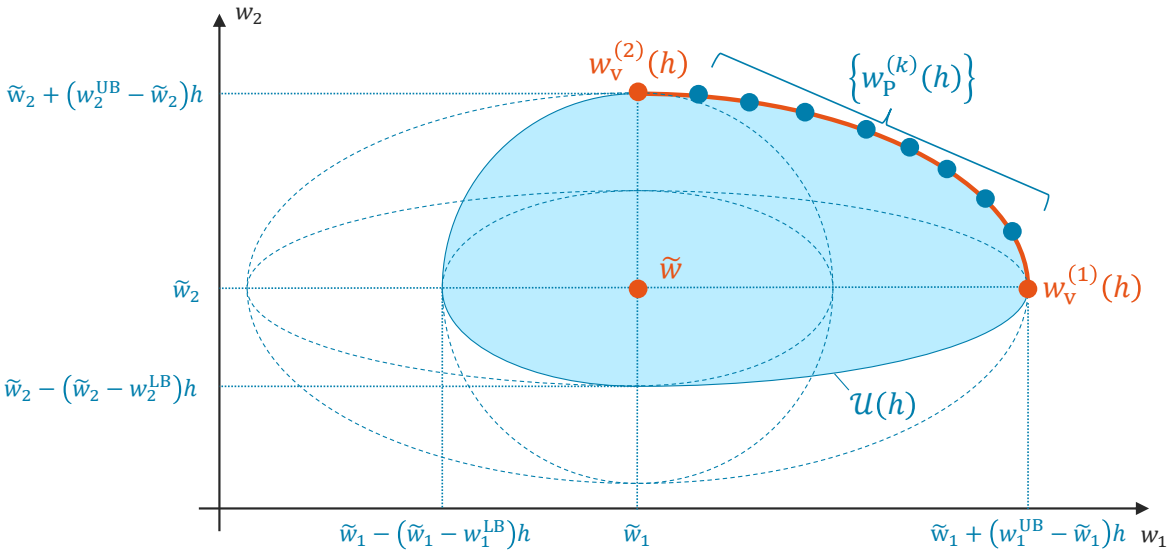


Figure 5.7 Examples of the uniform Pareto scenarios sampling and the vertices sampling. Here we assume scenario w has two uncertain parameters (w_1 and w_2), both of which have the direction of goodness in the negative direction (i.e., the lower, the better). When we sample scenarios to calculate the minimum cumulative reward in $\mathcal{U}(h)$, the Pareto vertices to be sampled will be $w_{\mathrm{v}}^{(1)}(h) = (\tilde{w}_1 + (w_1^{\mathrm{UB}} - \tilde{w}_1)h, \tilde{w}_2)$ and $w_{\mathrm{v}}^{(2)}(h) = (\tilde{w}_1, \tilde{w}_2 + (w_2^{\mathrm{UB}} - \tilde{w}_2)h)$, and the Pareto scenarios $\left\{w_{\mathrm{P}}^{(k)}(h)\right\}$ will be uniformly sampled from the Pareto front shown in orange.

5.4.2 HoU analysis results and discussion

The HoU plot is shown in Figure 5.8. Findings from the HoU plot are:

F.1 Under the nominal scenario ($h = 0$), **Aggressive** performs better than any other policy, **Staged (X, IR, G)** (three-staged) performs the worst, and the other two (**Staged (X/IR, G)** and **Staged (X, IR/G)**) perform approximately the same.

F.2 **Aggressive** performs the best in the best scenario, regardless of h.

F.3 The range of the cumulative reward under each policy gradually increases as h increases in $0 \leq h \leq 0.6$.

F.4 The difference between the four expert policies' worst-case performance becomes smaller as h increases in $0 \leq h \leq 0.6$.

F.5 The worst cumulative reward under each policy drastically decreases (let us call it a "drop") as h increases and reaches $h = 0.6$ to $h = 0.8$.

F.6 The magnitude of the "drop" is different under each expert policy. It is the largest under **Aggressive** and the smallest under **Staged (X, IR, G)** and under **Staged (X/IR, G)**.

F.1 is not surprising because the uncertain parameters were defined so that the development of `CORE3` is the fastest and the most inexpensive among the staged development strategies, as the project can optimize its resources and development process for the development of a single highly-integrated formation flying system capable of any of the three scientific observations. **F.2** suggests that **Aggressive**'s superiority is unchanged by the horizon of uncertainty if the decision-maker focuses on the best-case scenario. The gradual increase in the performance range mentioned in **F.3** is due to the uncertainty in each technology's development cost. The gradual increase is, in fact, linear in the horizon of uncertainty. This is because when all the missions are feasible at the final time step, and the eventually completed technologies

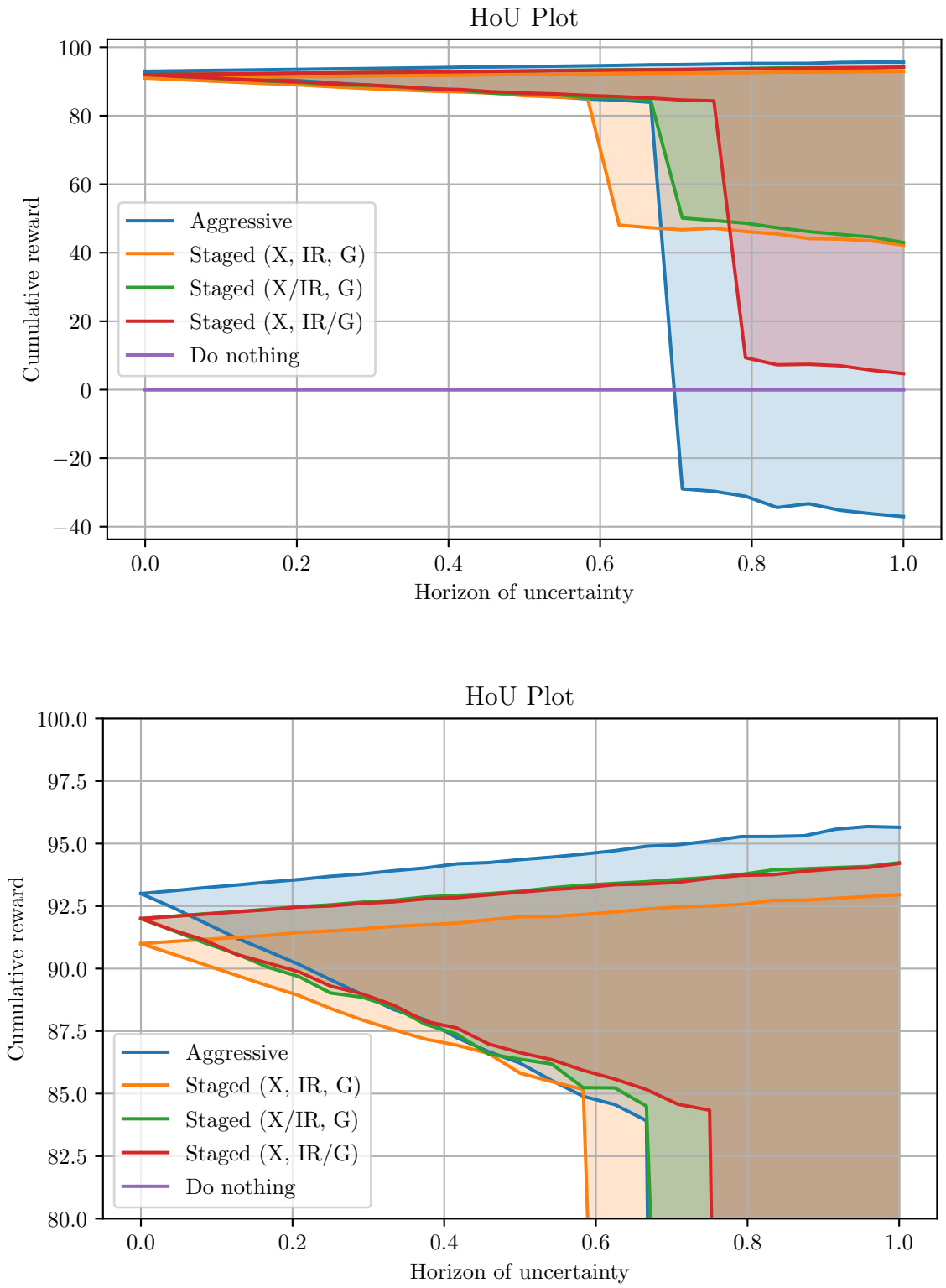


Figure 5.8 The HoU plot of the expert policies. The entire plot (up) and the zoomed plot (down).

τ are fixed, the cumulative reward is a linear combination of the deviation in the development cost of the completed technologies plus some constant. Therefore, the minimization of the cumulative reward in $w \in \mathcal{U}(h)$ is a minimization problem with a linear objective function and a quadratic constraint, and the optimal value changes linearly in h. Because the relative deviations in the development cost of technology upgrades (e.g., `CORE1to2`) are smaller than that of base technologies (e.g., `CORE1`), the increase in the performance range is smaller under staged policies, as observed in **F.4**. However, when h becomes so large that the worst (i.e., minimum) value of t_{lim} in $\mathcal{U}(h)$ is smaller than the time when all the missions become feasible, the agent can no longer receive the mission reward, and the worst-case cumulative reward drops drastically, as observed in **F.5**. The reason for **F.6** is that under **Aggressive**, none of the missions will become feasible when $h \geq 0.7$, whereas under **Staged (X, IR, G)** or **Staged (X/IR, G)**, both mission `X` and `IR` become feasible after the "drop," and under **Staged (X, IR/G)**, only mission `X` becomes feasible after the "drop."

5.4.3 Policy/HoU selection

Based on the HoU analysis results, the decision-maker selects policies and a horizon of uncertainty to consider in later analyses. In the `FormationFlying` problem, we selected the four expert policies as candidate policies and $h = 1$ as the horizon of uncertainty.

5.5 Scenario analysis

5.5.1 Scenario analysis settings

The parameters used in the scenario analysis are shown in Table 5.6. 10,000 scenarios were sampled quasi-uniformly from $\mathcal{U}(1)$. Under each scenario, the four candidate policies were simulated, resulting in 40,000 simulations in total. The performance measures for each simulation were: the cumulative reward, whether the three missions

Table 5.6 Parameters used in the scenario analysis of the `FormationFlying` problem.

Parameter	Value
Discount rate γ	1
Scenarios sampling method	Uniform scenarios sampling
The number of scenarios	10,000

become feasible before the time limit, and when the three missions become feasible if they do.

The scenarios were quasi-uniformly sampled from $\mathcal{U}(1)$ as follows. Let n_{w} be the number of scenarios to sample. First, sample uniformly-distributed n_{w} random points inside a d_{w}-dimensional unit ball by the following steps:

Step 1. Sample an $n_{\mathrm{w}} \times d_{\mathrm{w}}$ matrix $\mathbf{A}$ whose elements are independently and identically distributed as $a_{ij} \sim \mathcal{N}(0, 1)$.

Step 2. Sample an $n_{\mathrm{w}} \times d_{\mathrm{w}}$ matrix $\mathbf{B}$ whose elements are independently and identically distributed as $b_{ij} \sim U(0, 1)$.

Step 3. Construct an $n_{\mathrm{w}} \times d_{\mathrm{w}}$ matrix $\mathbf{X}$ with $x_{ij} \equiv \dfrac{a_{ij}}{\|\mathbf{a}^{(i)}\|} b_{ij}^{\frac{1}{d_{\mathrm{w}}}}$ where $\mathbf{a}^{(i)} \equiv \begin{bmatrix} a_{i1} & \dots & a_{id_{\mathrm{w}}} \end{bmatrix}$. The row vectors of X are uniformly distributed inside a d_{w}-dimensional unit ball.

For each point $\mathbf{x}^{(i)}$, x_{ij} is allocated to the j-th uncertain parameter as its "budget of uncertainty." If $x_{ij} = 0$, the uncertain parameter will be set to its nominal value as $w_j = \tilde{w}_j$. Otherwise, there are two w_j's that satisfy $d_j(w_j) = |x_{ij}|$. Let us denote the two solutions as $w_j^{(1)}$ and $w_j^{(2)}$, and assume without the loss of generality that $w_j^{(1)}$ is better than $w_j^{(2)}$. For example, $w_j^{(1)} < w_j^{(2)}$ if w_j is a cost. The uncertain parameter will then be set to the better value if $x_{ij} > 0$ or the worse value if $x_{ij} < 0$.

The sampling of the scenarios is *quasi*-uniform in that it is *not* uniform with respect to the space $\mathcal{U}(h)$, but rather with respect to the value of each uncertain parameter's distance function.

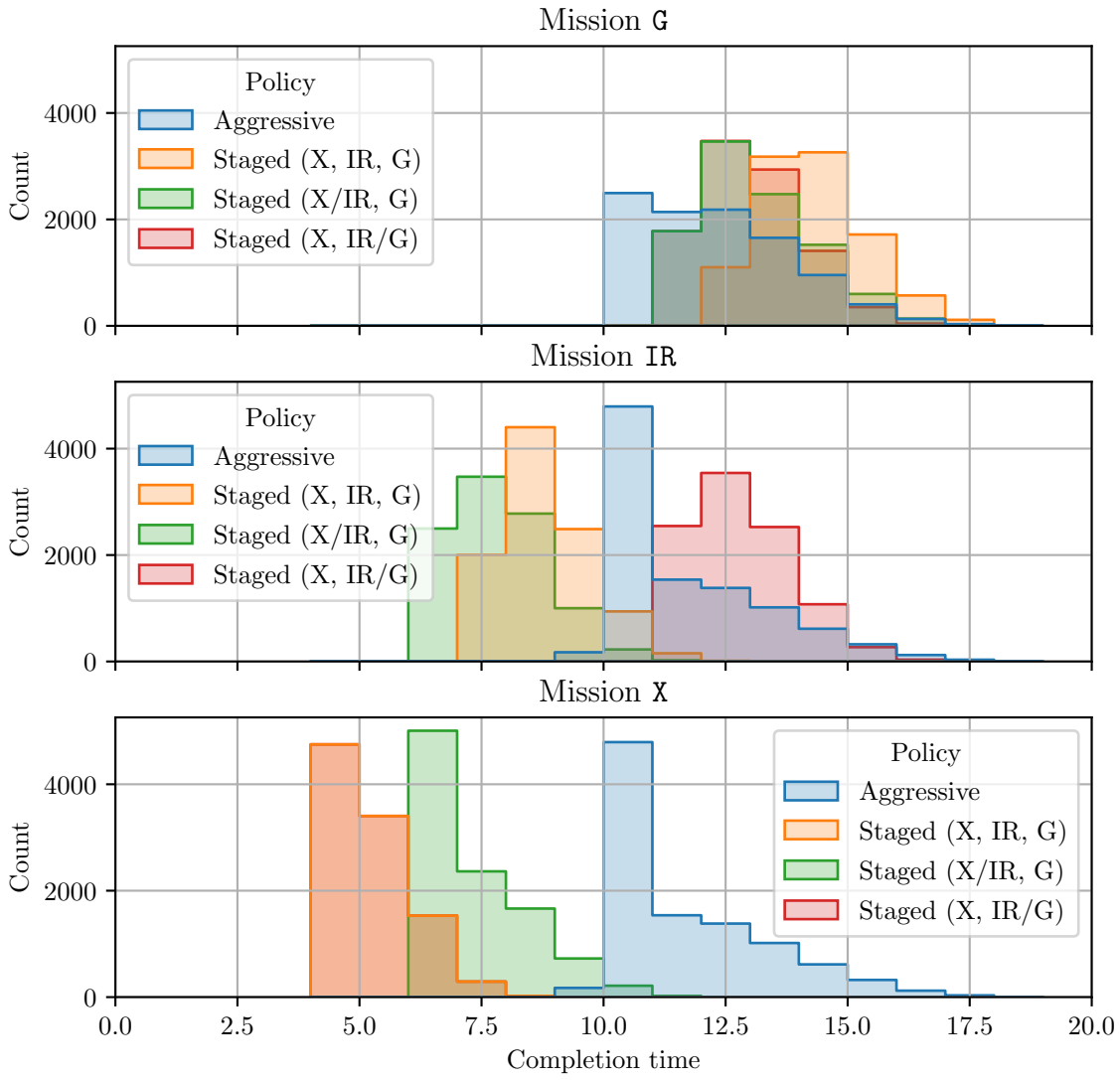
5.5.2 Scenario analysis results and discussion

Figures 5.9 and 5.10 shows the distribution of when each mission becomes feasible under each expert policy, under the 10,000 sampled scenarios. It can be seen that the distributions of the completion time are right-tailed. It should also be noted that the difference in the completion time for mission `G` is smaller than the difference in the completion time for mission `X` or `IR`. If the decision-maker has a motivation to conduct any of the missions sooner, staged development policies may be preferred. On the other hand, if the time limit's uncertainty is small, the aggressive development policy is the best choice.

We analyzed each uncertain parameter's sensitivity against each performance measure by calculating the feature score in the regression using the extremely randomized trees [79]. The scores under each policy are shown in Figure 5.11. Under **Aggressive**, **Staged (X/IR, G)**, and **Staged (X, IR/G)**, the development costs of the developed technologies have high sensitivity against the cumulative reward, whereas, under **Staged (X, IR, G)**, the time limit has the highest sensitivity. This indicates that **Staged (X, IR, G)** is more vulnerable to the uncertainty in the time limit than the other policies.

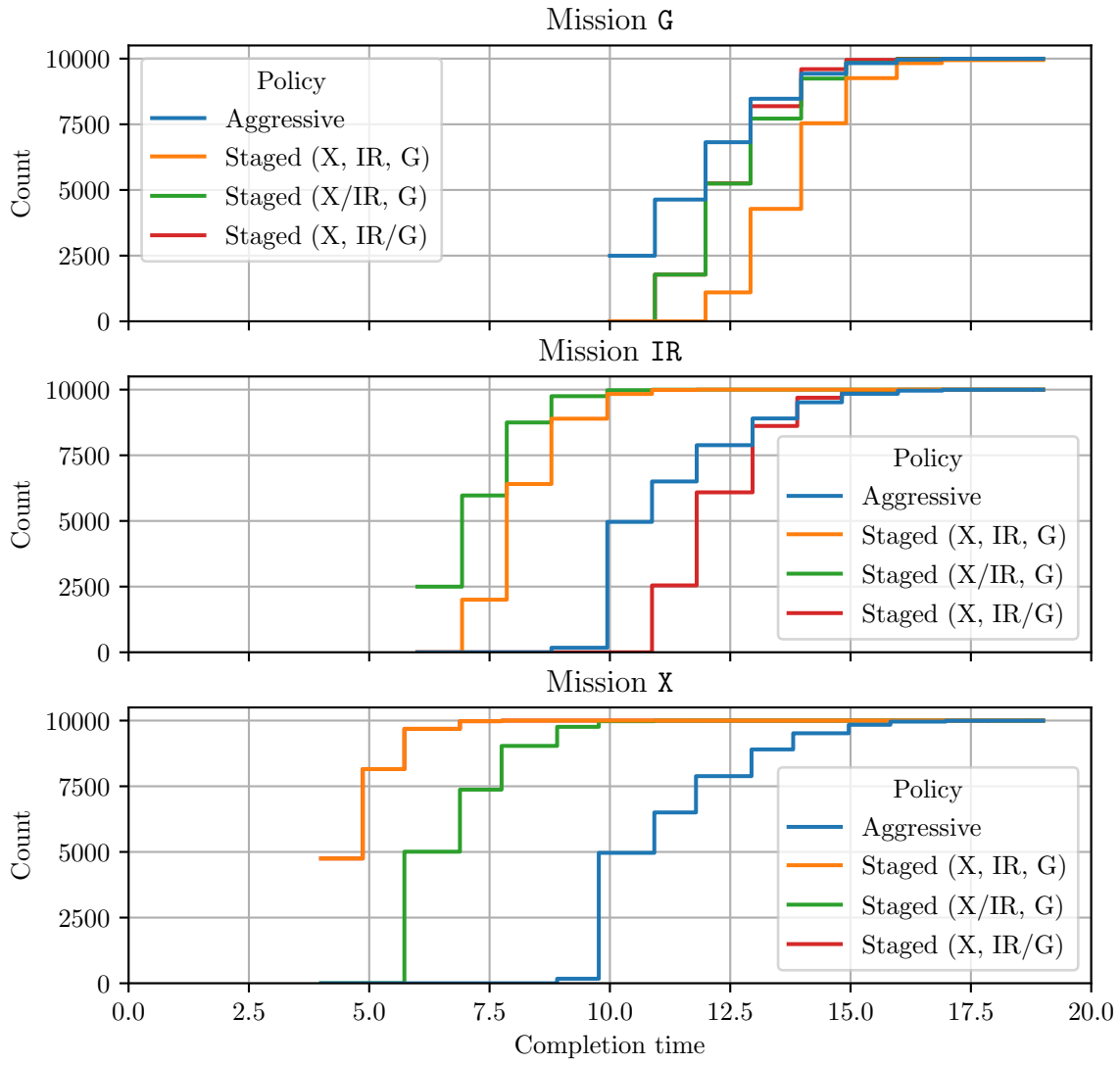
We conducted the scenario discovery to the simulation results by defining the cases of interest as the simulation cases where all the missions became feasible by the time limit. The Pareto front in the density–coverage space obtained with the PRIM under each policy is shown in Figure 5.12. The box with the maximum density and the one with the maximum coverage are close to each other under every policy, indicating the cases of interest and the other cases are separated in the scenario space. Also, the number of restricted dimensions were 1 at maximum.

Figure 5.13 shows the distribution of the cases of interest and the other cases in the restricted dimension of the box with the largest density. The development cost of `CORE3` was the restricted dimension under **Aggressive**, where the time limit was under the other three policies. This is because `CORE3` is the only technology whose



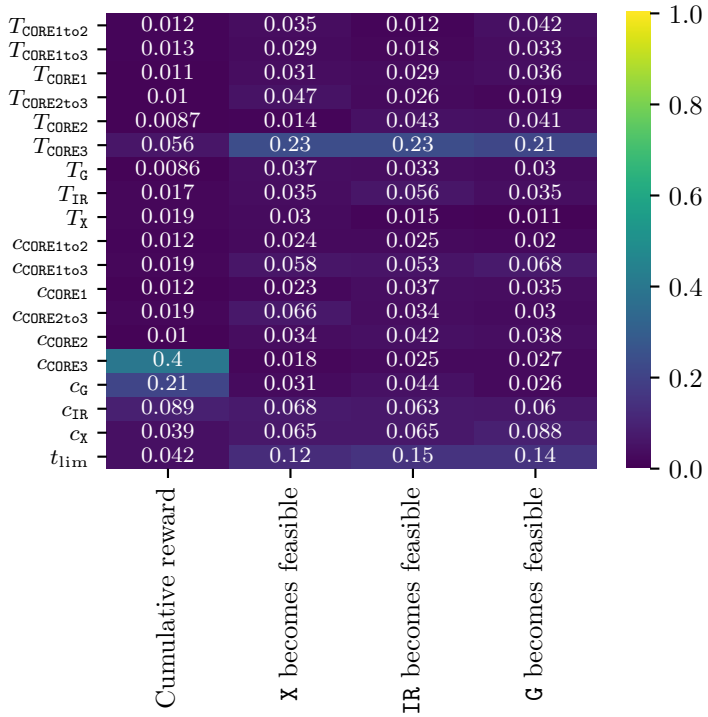
(a) The histogram plot

Figure 5.9 Distribution of the time when each mission becomes feasible under each expert policy, under scenarios in $\mathcal{U}(1)$. Note that the distribution of the completion time of mission `X` is identical under **Staged (X, IR, G)** and **Staged (X, IR/G)** and thus cannot be distinguished in both plots.

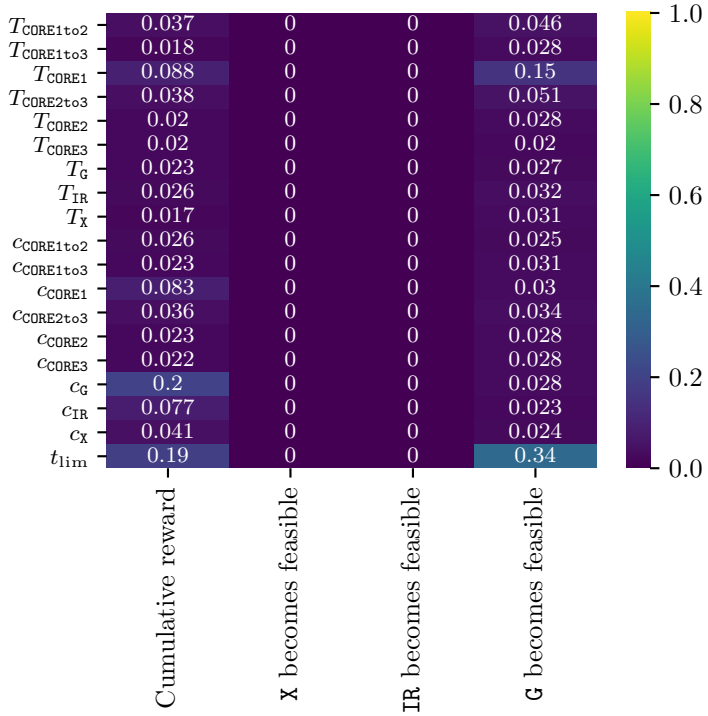


(a) The cumulative plot

Figure 5.10 Distribution of the time when each mission becomes feasible under each expert policy, under scenarios in $\mathcal{U}(1)$ (cont.). Note that the distribution of the completion time of mission `X` is identical under **Staged (X, IR, G)** and **Staged (X, IR/G)** and thus cannot be distinguished in both plots.

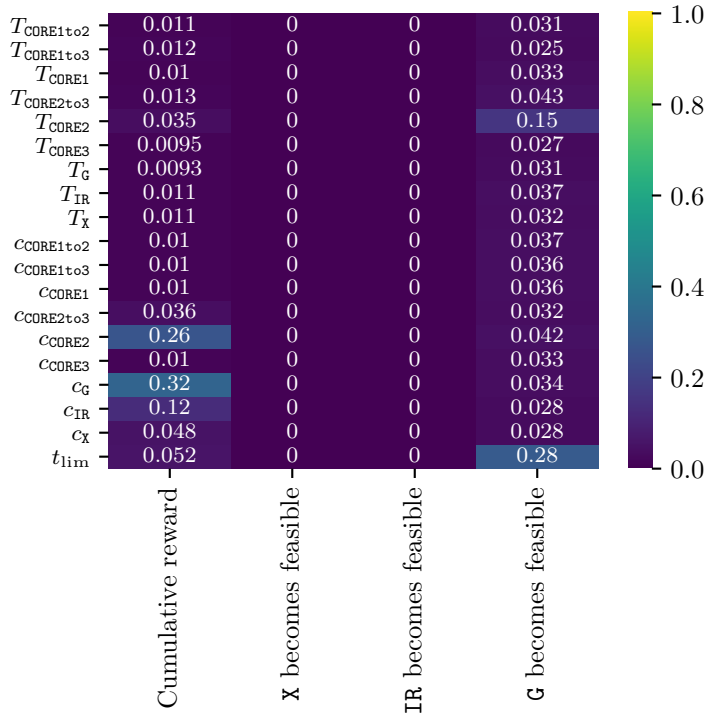


(a) Policy: Aggressive

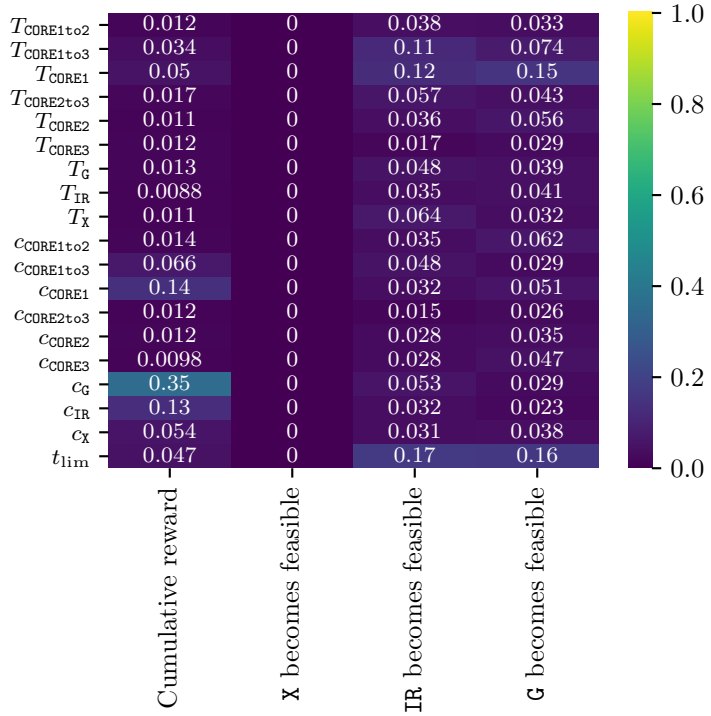


(b) Policy: Staged (X, IR, G)

Figure 5.11 Feature scoring of each uncertain parameter under each policy.

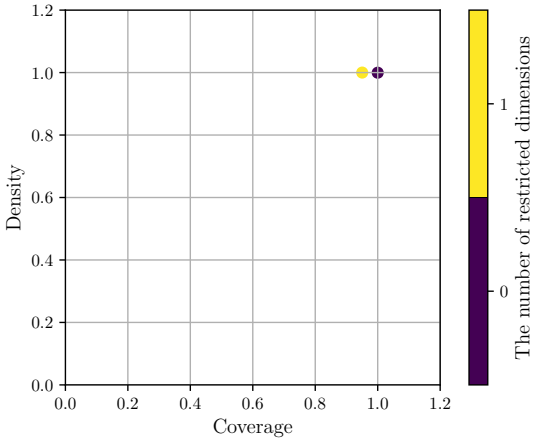


(c) Policy: Staged (X/IR, G)

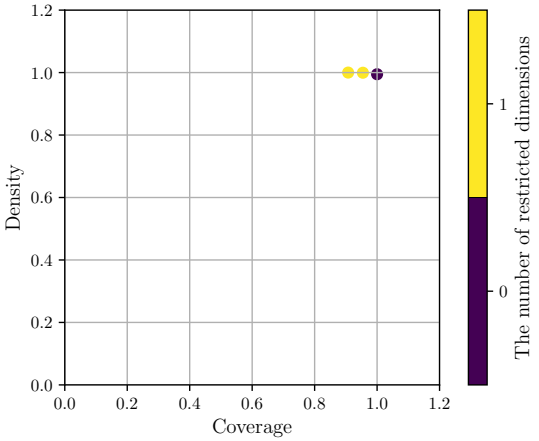


(d) Policy: Staged (X, IR/G)

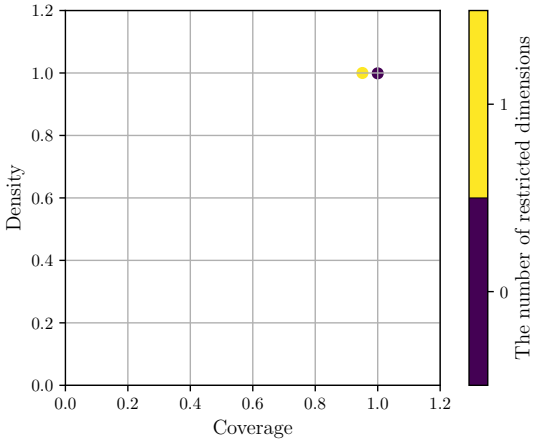
Figure 5.11 Feature scoring of each uncertain parameter under each policy (cont.).



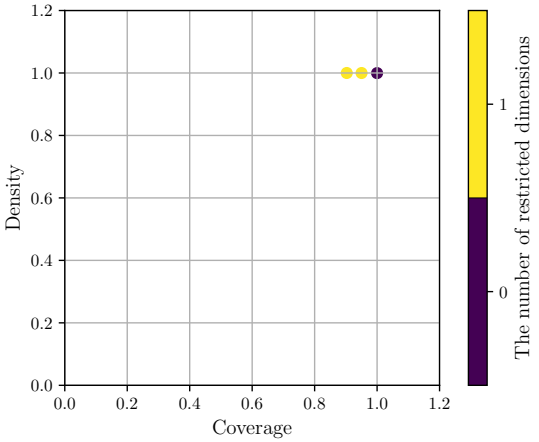
(a) Policy: Aggressive



(b) Policy: Staged (X, IR, G)



(c) Policy: Staged (X/IR, G)



(d) Policy: Staged (X, IR/G)

Figure 5.12 Density–coverage Pareto front of each policy. The cases of interest were defined as the ones where all of the three missions are conductible at the final time step.

development cost has an upper deviation larger than the lower deviation of the time limit, i.e., $T^{\mathrm{UB}}_{\texttt{CORE3}} - \tilde{T}_{\texttt{CORE3}} > \tilde{t}_{\mathrm{lim}} - t^{\mathrm{LB}}_{\mathrm{lim}}$.

Figure 5.14 shows the regional sensitivity to observe each uncertain parameter's sensitivity against whether all the missions became feasible by the time limit. It can be seen that the development cost of each technology has no sensitivity. The uncertain parameters that have high sensitivity according to the plot are: `CORE3`'s development time and the time limit under **Aggressive**, `CORE1`'s development time and the time limit under **Staged (X, IR, G)**, `CORE2`'s development time and the time limit under **Staged (X/IR, G)**, and `CORE1`'s development time, `CORE1to3`'s development time, and the time limit under **Staged (X, IR/G)**.

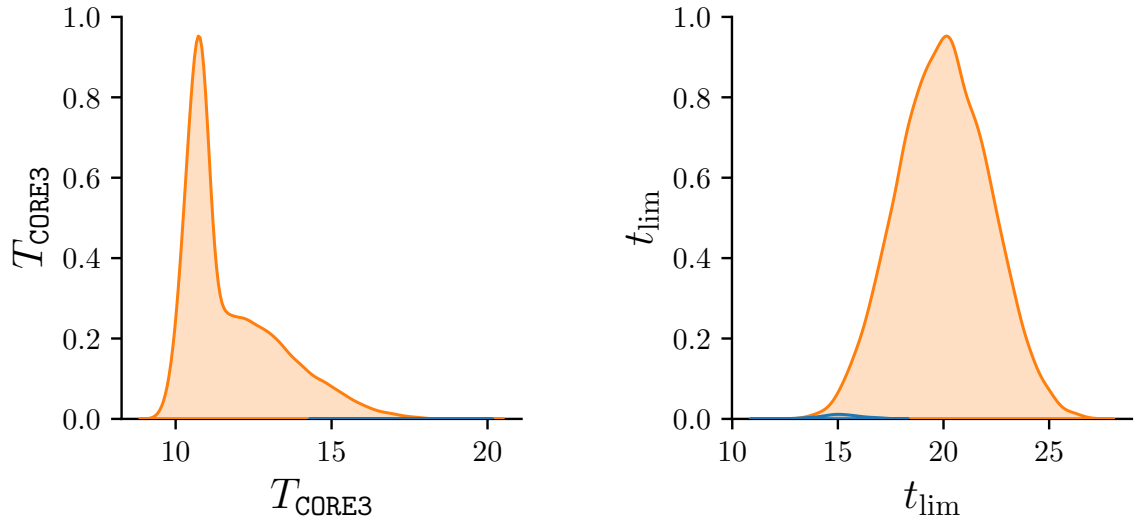
5.6 Discussion

Table 5.7 summarizes each policy's advantages and disadvantages based on the findings from the HoU analysis and the scenario analysis. The final decision is up to the decision-maker's attitude toward risk.

If the decision-maker is interested more in the best-case scenario than in the worst-case, **Aggressive** may be the choice because there is a possibility that the low total development costs compared to the other policies.

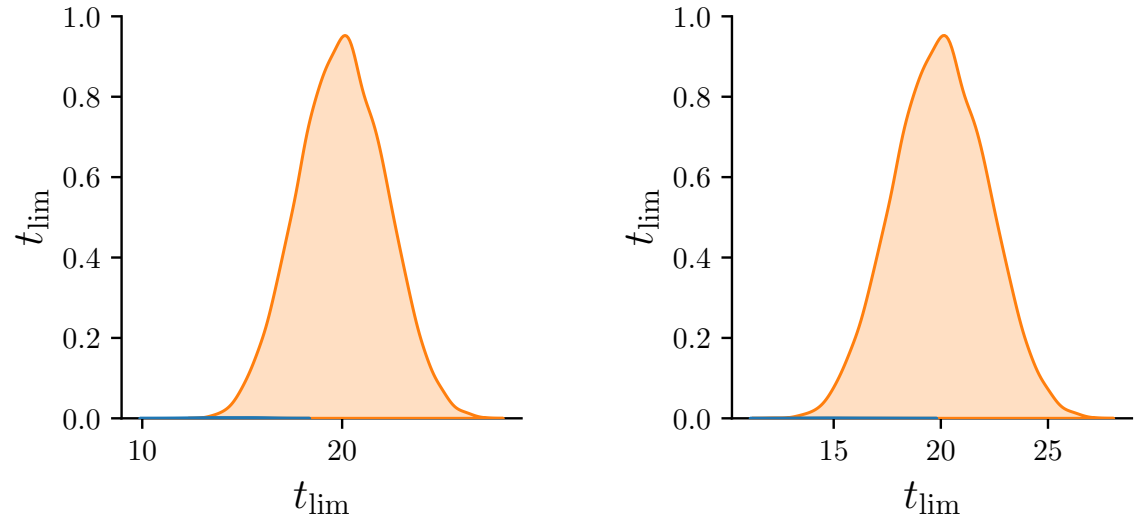
If the decision-maker considers the uncertainty in the time limit to be large and wants missions to be feasible as soon as possible, but allows mission `G` not to be conducted by the time limit, **Staged (X, IR, G)** may be the choice because one can conduct each mission as soon as the level of the core formation flying technology reaches the required level.

If the decision-maker considers the uncertainty in the time limit to be large and wants to conduct all the missions, **Staged (X/IR, G)** may be the choice because only mission `G` has a possibility of not becoming feasible by the time limit, and the possibility is lower than under **Staged (X, IR, G)**.



(a) Policy: Aggressive

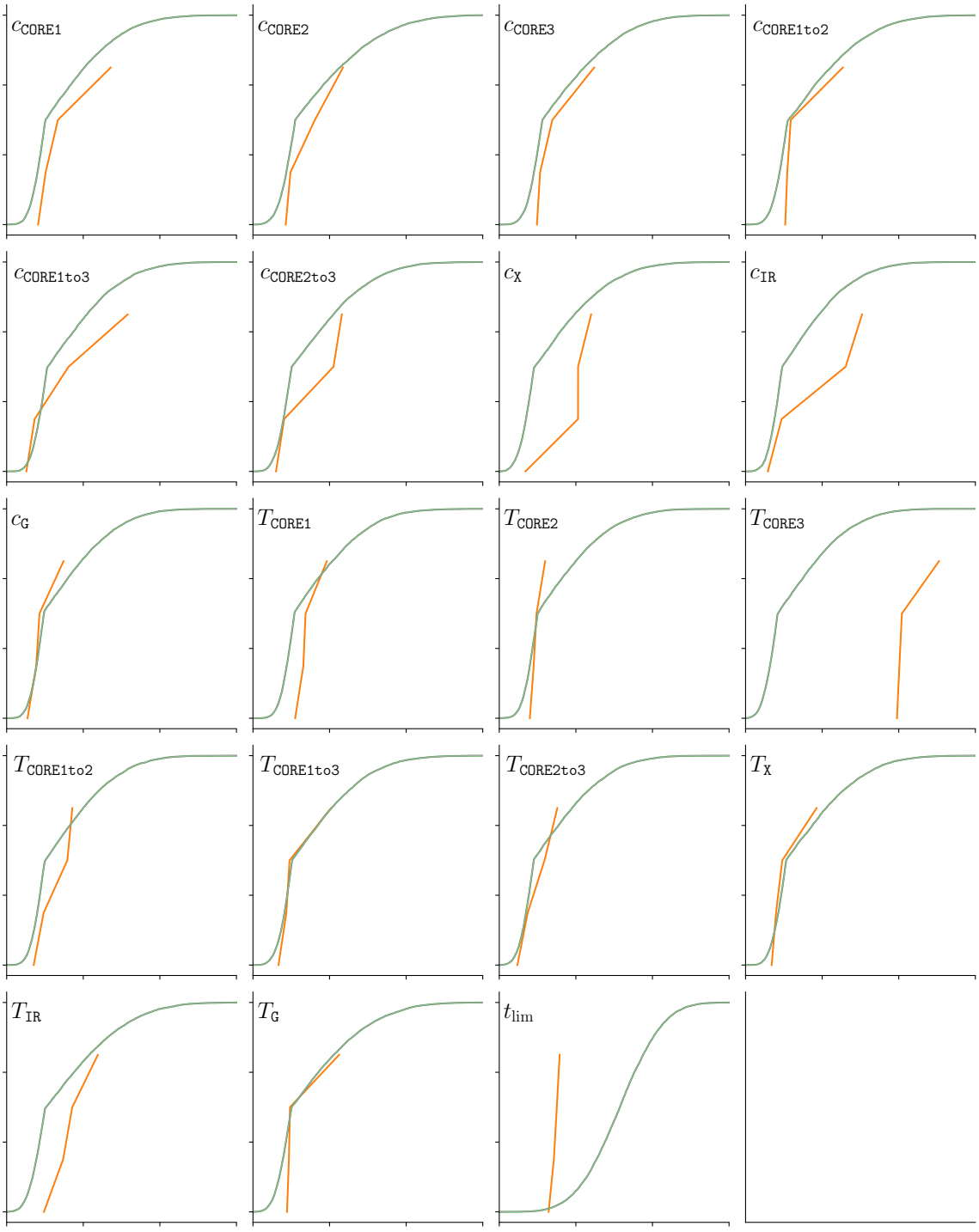
(b) Policy: Staged (X, IR, G)



(c) Policy: Staged (X/IR, G)

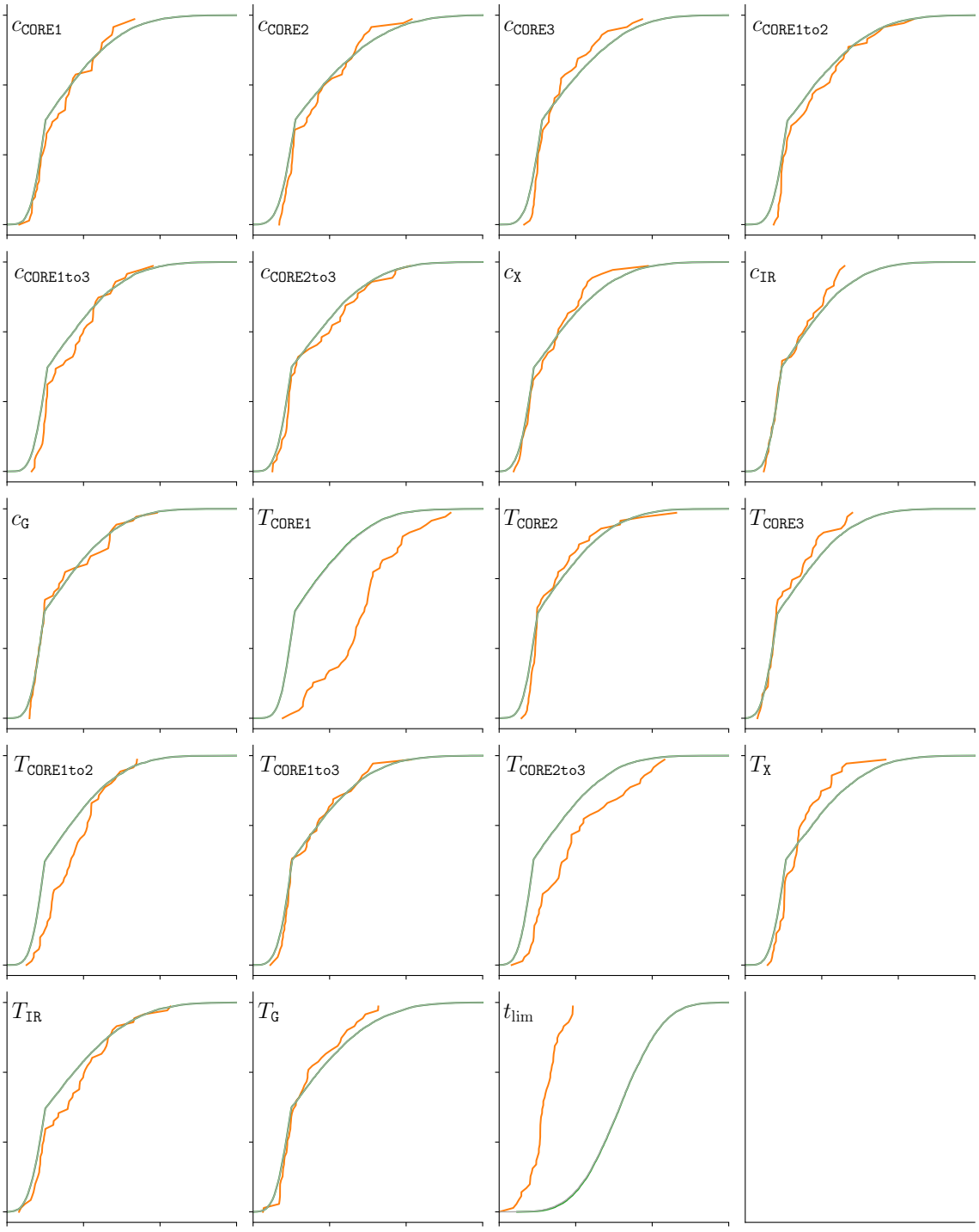
(d) Policy: Staged (X, IR/G)

Figure 5.13 The distribution of the cases of interest (orange) and the other cases (blue) in the restricted dimension.



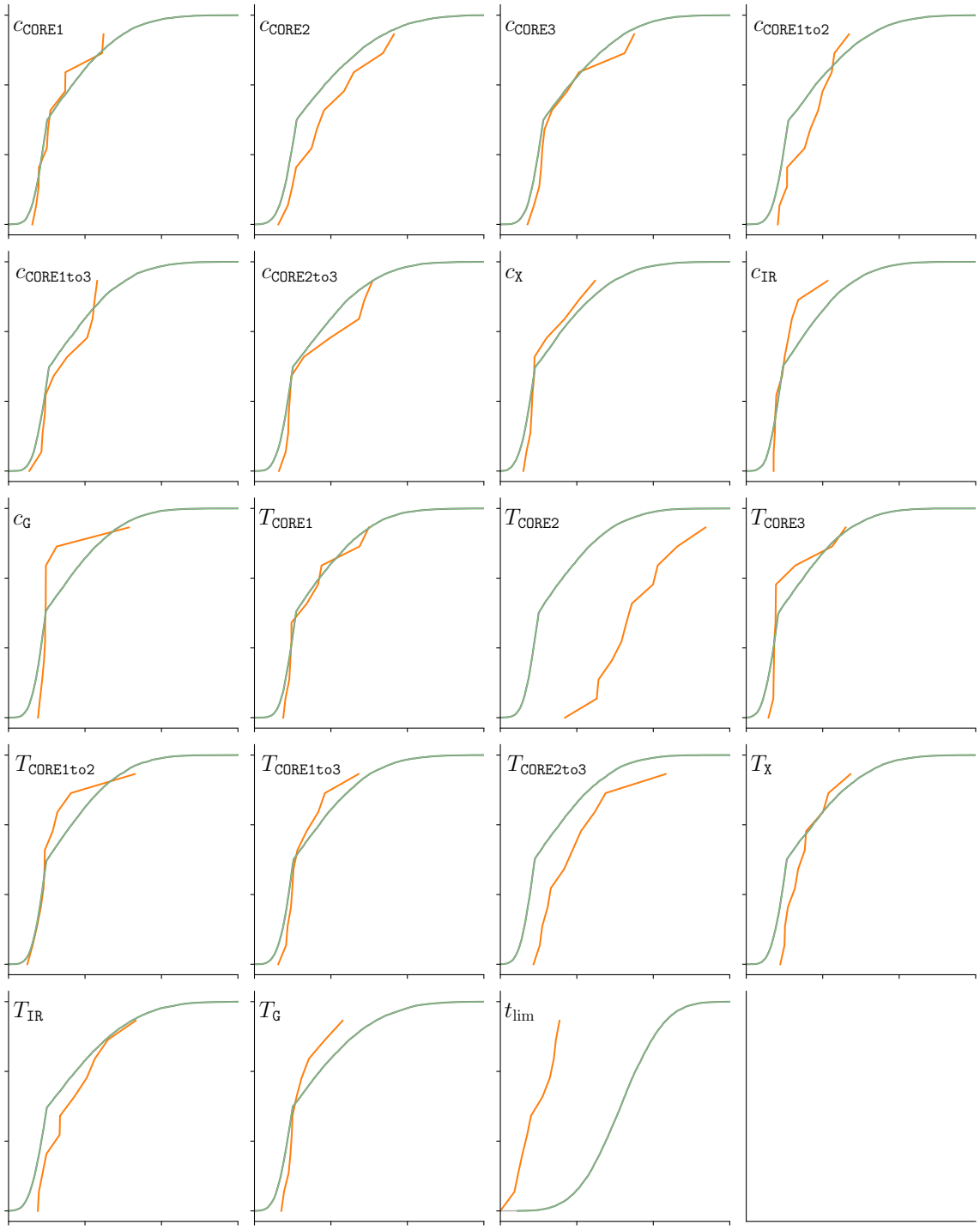
(a) Policy: Aggressive

Figure 5.14 The regional sensitivity analysis of the cases of interest under each policy. The green curve is the cumulative plot of the cases of interest projected onto the uncertain parameter, and the orange curve is that of the other cases.



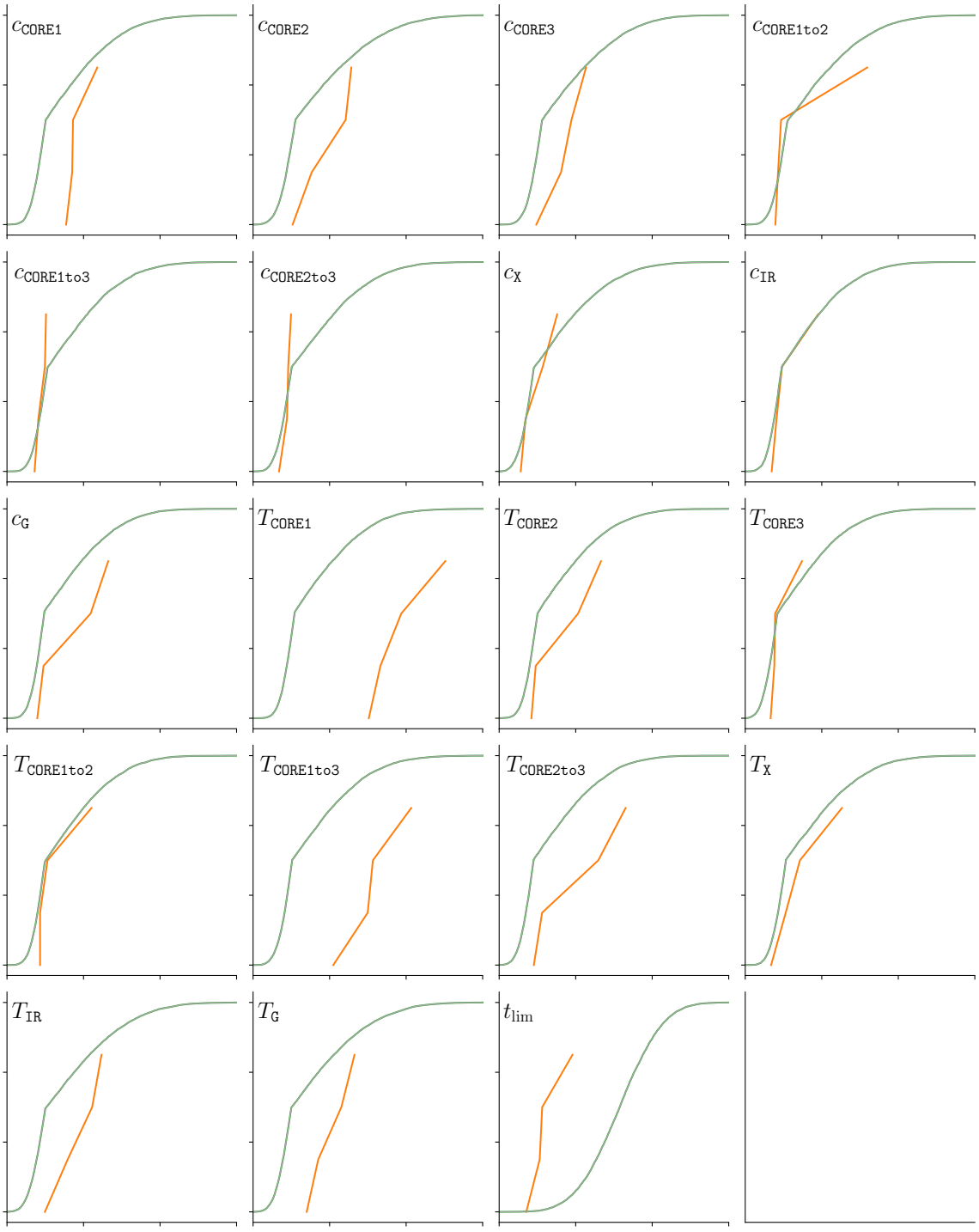
(b) Policy: Staged (X, IR, G)

Figure 5.14 The regional sensitivity analysis of the cases of interest under each policy. The green curve is the cumulative plot of the cases of interest projected onto the uncertain parameter, and the orange curve is that of the other cases (cont.).



(c) Policy: Staged (X/IR, G)

Figure 5.14 The regional sensitivity analysis of the cases of interest under each policy. The green curve is the cumulative plot of the cases of interest projected onto the uncertain parameter, and the orange curve is that of the other cases (cont.).



(d) Policy: Staged (X, IR/G)

Figure 5.14 The regional sensitivity analysis of the cases of interest under each policy. The green curve is the cumulative plot of the cases of interest projected onto the uncertain parameter, and the orange curve is that of the other cases (cont.).

Finally, if the decision-maker cares whether all the missions become feasible by the time limit and does not value cases where only some of the missions become feasible, **Staged (X, IR/G)** may be the choice because all the missions become feasible in the largest range of h.

5.7 Expert feedback

We shared the results with an expert working in a formation flying mission project and obtained the following feedback:

Benefits from the results

- The framework helps quantitatively recognize the problem with its uncertainty.
- While the results do not contradict with the intuition, unlike the intuition, it is good to see the results visually and quantitatively.

Opportunities for enhancement

- It will be interesting to see the results under a different definition of the uncertainty model with different lower and upper bound of each uncertain parameter.
- Creating the uncertainty model is difficult because the uncertainty is too large even for an expert to estimate each parameter's uncertainty.

It can be safely said that the results from the MSRDM framework can benefit an actual decision-making problem, though there still exist some possible enhancements to be made, especially in the creation of the uncertainty model.

Table 5.7 Advantages and disadvantages of each policy.

Policy	Advantages	Disadvantages
Aggressive	• The cumulative reward is the highest in the nominal and best-case scenarios.	• There is a risk of no mission being conducted due to the long development time of `CORE3`.
Staged (X, IR, G)	• The effect of the time limit uncertainty is limited to mission `G`. • The completion time of mission `X` and `IR` are among the shortest.	• The time limit has the largest sensitivity due to slow development. • The likelihood of mission `G` not being conducted is larger than the other policies. • The cumulative reward is the lowest in the nominal and the best-case scenarios.
Staged (X/IR, G)	• The effect of the time limit uncertainty is limited to mission `G`. • The completion time of mission `IR` is the shortest.	• Mission `G` may not be conducted due to the early time limit.
Staged (X, IR/G)	• All the missions are conducted in the largest range of h.	• There is a risk that only mission `X` is conducted due to the long development time of `CORE1` and the early time limit.

Chapter 6

Case Study II: Technology Roadmapping of Marine Propulsion System

6.1 Background

Third IMO GHG Study 2014 [95] reported that international shipping, carrying as much as 90 % of the world trade by volume, emitted approximately 2.2 % of the total emission volume in 2012. It also forecasted that CO_2 emissions from international shipping could increase by 50 % to 250 % by 2050. To address GHG emissions from international shipping, the International Maritime Organization (IMO) published "Initial IMO Strategy on Reduction of GHG Emissions from Ships [96]" in 2018, where it shared its three levels of ambition. The level-one ambition is for "carbon intensity of the ship to decline through implementation of further phases of the energy efficiency design index (EEDI) for new ships." The level-two ambition is "to reduce CO_2 emissions per transport work, as an average across international shipping, by at least 40 % by 2030, pursuing efforts towards 70 % by 2050, compared to 2008." The level-three ambition is "to peak GHG emissions from international shipping as soon as possible

and to reduce the total annual GHG emissions by at least 50 % by 2050 compared to 2008."

The report by The Japan Ship Technology Research Association and The Ministry of Land, Infrastructure, Transport and Tourism of Japan [97] considers two possibilities of future pathways to reduce the maritime CO_2 emission: "a fuel shift from LNG to carbon-recycled methane" and "the expansion of hydrogen and/or ammonia fuels." The former pathway assumes that LNG ships and infrastructure to supply the fuel will be commonplace while infrastructure to supply hydrogen or ammonia will not be as available. The latter assumes that supply chain of hydrogen and ammonia will be ubiquitous. ICEs with ammonia [98, 99] and hydrogen [100, 101] are both studied for maritime use, and hydrogen and ammonia are regarded as alternative fuels in the medium and long term by IMO along with biofuels [102].

6.2 Decision structuring

6.2.1 Identifying relevant parameters using the XLRM framework

The external factors, policy levers, and performance metrics of the `MarinePropulsion` problem were defined as shown in Table 6.1.

We identified the external factors, i.e., the uncertainties, the development time and cost of each technology, the fuel scenario, and the time of the fuel scenario reveal. The uncertainty in the technology development derives from limited estimation capability, unexpected effort due to technical issues during the development process, and schedule slip. A strategy is also affected by how widely the infrastructure required for each propulsion configuration is spread globally. Therefore, we defined the fuel scenario (which of the two pathways will be realized) and when the fuel scenario becomes known to the decision-maker as external factors.

Table 6.1 List of external factors, policy levers, and performance metrics in the `MarinePropulsion` problem.

External factors (X)	Development cost of each technology Development time of each technology Fuel scenario Time of the fuel scenario reveal
Policy levers (L)	Which configuration to develop
Performance metrics (M)	Achievement of the two IMO CO_2 reduction goals

The policy lever in the `MarinePropulsion` problem is the selection of propulsion configuration to develop. Table 6.2a lists the candidate marine propulsion configurations for reducing CO_2 emission. One of the options is the CO_2 collection. Although it is not a propulsion system, it can collect the CO_2 contained in propulsion system's emission and convert it into methane, which can be reused as fuel. The other four options are internal combustion engines (ICE). The first ICE option is an ICE with mixed fuel of ammonia and heavy oil.

6.2.2 Defining the technologies and configurations

Table 6.2a shows the five candidate configurations that reduce the CO_2 emission from the marine propulsion: `A&HO`, `A`, `H&M`, `H`, and `CC`. `A&HO` is the ICE with mixed fuel of ammonia and heavy oil, `A` is the ICE with mono fuel of ammonia, `H&M` is the ICE with mixed fuel of hydrogen and methane, `H` is the ICE with mono fuel of hydrogen, and `CC` is the CO_2 collection. For each configuration to be built, some technologies need to be developed. As shown in Table 6.2b, five technologies were identified: `A&HO`, `AtoMONO`, `H&M`, `HtoMONO`, and `CC`. `A&HO` is the technology required for the ICE with mixed fuel of ammonia and heavy oil (configuration `A&HO`), `H&M` is the technology required for the ICE with mixed fuel of hydrogen and methane (configuration `H&M`), and `CC` is the technology required for the CO_2 collection (configuration `CC`). Technologies `AtoMONO` and `HtoMONO` are upgrades from mixed-fuel combustion (`A&HO` and `H&M`, respectively) to mono-fuel combustion.

Table 6.2 Technologies and configurations in the `MarinePropulsion` problem.

(a) Configurations

Name	Description	Required technologies
`A&HO`	ICE with mixed fuel of ammonia and heavy oil	`A&HO`
`A`	ICE with mono fuel of ammonia	`A&HO`, `AtoMONO`
`H&M`	ICE with mixed fuel of hydrogen and methane	`H&M`
`H`	ICE with mono fuel of hydrogen	`H&M`, `AtoMONO`
`CC`	CO_2 collection	`CC`

(b) Technologies

Name	Description
`A&HO`	ICE with mixed fuel of ammonia and heavy oil
`AtoMONO`	Upgrade from `A&HO` to ICE with mono fuel of ammonia
`H&M`	ICE with mixed fuel of hydrogen and methane
`HtoMONO`	Upgrade from `H&M` to ICE with mono fuel of hydrogen
`CC`	CO_2 collection

Let n the number of technologies and m the number of configurations. Note that the technology names (`A&HO`, `AtoMONO`, …) and their indices (1, 2, …) are used interchangeably, and so are the configuration names (`A&HO`, `A`, …) and their indices (1, 2, …). We defined the configuration feasibility function $f_{\text{feas},j}(\tau)\colon 2^{\{1,\dots,n\}} \to \{\texttt{True}, \texttt{False}\}$ ($j = 1, \dots, m$) representing whether the configuration j can be built with the set of technologies τ as:

$$f_{\text{feas,A\&HO}}(\tau) = \texttt{A\&HO} \in \tau \tag{6.1a}$$

$$f_{\text{feas,A}}(\tau) = \{\texttt{A\&HO}, \texttt{AtoMONO}\} \subseteq \tau \tag{6.1b}$$

$$f_{\text{feas,H\&M}}(\tau) = \texttt{H\&M} \in \tau \tag{6.1c}$$

$$f_{\text{feas,H}}(\tau) = \{\texttt{H\&M}, \texttt{HtoMONO}\} \subseteq \tau \tag{6.1d}$$

$$f_{\text{feas,CC}}(\tau) = \texttt{CC} \in \tau \tag{6.1e}$$

For example, the definition of $f_{\text{feas,A}}(\tau)$ indicates that configuration `A` can be built if `A&HO` and `AtoMONO` are developed.

6.2.3 Defining the non-probabilistic uncertainty model

The `MarinePropulsion` problem has 17 uncertain parameters: the fuel scenario $\phi \in \{\texttt{CRM}, \texttt{HA}\}$, the time of the fuel scenario reveal $t_{\text{rev}} \in \mathbb{R}$, the development cost $c_i \in \mathbb{R}$ and development time $T_i \in \mathbb{R}$ of each of the five technologies, and the CO_2 emission reduction performance η_j of each of the five configurations. Here we denote scenario w as

$$w = (\phi, t_{\text{rev}}, c_1, \ldots, c_n, T_1, \ldots, T_n, \eta_1, \ldots, \eta_m) \equiv (w_1, \ldots, w_{d_{\text{w}}}) \tag{6.2}$$

where $d_{\text{w}} = 2 + 2n + m = 17$ is the number of uncertain parameters (i.e., the dimension of w). The uncertainty model $\mathcal{U}(h)$ $(0 \leq h \leq 1)$ was defined as an ellipsoid. Formally, the uncertainty model was defined as:

$$\mathcal{U}(h) \equiv \left\{ (w_1, \ldots, w_{d_{\text{w}}}) \;\middle|\; \sum_{i=1}^{d_{\text{w}}} \left(d_i(w_i)\right)^2 \leq h^2 \right\}, \; 0 \leq h \leq 1 \tag{6.3}$$

where $d_i(w_i)$ is the distance function that represents the "distance" of the value w_i from the nominal value $\tilde{w}_i$:

$$d_1(\phi) = \begin{cases} d_{\texttt{CRM}} & (\text{if } \phi = \texttt{CRM}) \\ d_{\texttt{HA}} & (\text{if } \phi = \texttt{HA}) \end{cases} \tag{6.4a}$$

$$2 \leq i \leq d_{\text{w}}\colon \;\; d_i(w_i) = \frac{w_i - \tilde{w}_i}{\Delta w_i} \tag{6.4b}$$

$d_{\texttt{CRM}}$ and $d_{\texttt{HA}}$ are predefined constants in $[0, 1]$, representing how likely each discrete scenario is. Note that $d_{\texttt{CRM}} = 0$ or $d_{\texttt{HA}} = 0$ because otherwise neither fuel scenario would be considered possible in the nominal scenario, i.e., $\mathcal{U}(0) = \emptyset$. Δw_i is the maximum deviation of w_i from the nominal value. The parameter values are shown in Table 6.3.

6.2.4 Defining the problem as an MSRDM-MDP

We defined an MSRDM-MDP of the `MarinePropulsion` problem as follows.

Table 6.3 Definitions of the uncertain parameters in the `FormationFlying` problem.

(a) The distance of each fuel scenario.

Parameter	Value
$d_{\texttt{CRM}}$	0
$d_{\texttt{HA}}$	0

(b) The nominal value, the maximum deviation, the lower bound, and the upper bound of the time of the fuel scenario reveal t_{rev}.

Nominal $\tilde{t}_{\text{rev}}$	Maximum deviation Δt_{rev}	Lower bound	Upper bound
2027.5	2.5	2025	2030

(c) The nominal value, the lower bound, and the upper bound of the development cost of each technology c_i. The percentage represents the deviation rate from the nominal value.

Technology	Nominal	Lower bound		Upper bound	
`A&HO`	18.0	7.0	(−61 %)	29.0	(+61 %)
`AtoMONO`	18.0	7.0	(−61 %)	29.0	(+61 %)
`H&M`	26.0	10.0	(−62 %)	42.0	(+62 %)
`HtoMONO`	27.0	11.0	(−59 %)	43.0	(+59 %)
`CC`	14.0	3.0	(−79 %)	25.0	(+79 %)

(d) The nominal value, the lower bound, and the upper bound of the development time of each technology T_i. The percentage represents the deviation rate from the nominal value.

Technology	Nominal	Lower bound		Upper bound	
`A&HO`	5.5	3.0	(−45 %)	8.0	(+45 %)
`AtoMONO`	3.5	2.0	(−43 %)	5.0	(+43 %)
`H&M`	5.5	3.0	(−45 %)	8.0	(+45 %)
`HtoMONO`	3.5	2.0	(−43 %)	5.0	(+43 %)
`CC`	9.0	4.0	(−56 %)	14.0	(+56 %)

(e) The nominal value, the lower bound, and the upper bound of each configuration's CO_2 emission reduction η_j. The percentage represents the deviation rate from the nominal value.

Configuration	Nominal	Lower bound		Upper bound	
`A&HO`	0.55	0.4	(−27 %)	7.0	(+27 %)
`A`	0.94	0.9	(−4 %)	0.98	(+4 %)
`H&M`	0.45	0.3	(−33 %)	0.6	(+33 %)
`H`	0.95	0.9	(−5 %)	1.0	(+5 %)
`CC`	0.625	0.3	(−52 %)	0.95	(+52 %)

We defined a state as a vector $s = (t, u_1, \ldots, u_n, v_1, \ldots, v_n)$ where t is the current time, u_i is technology i's time under development, and v_i is a Boolean flag indicating whether the development of technology i is completed.

As shown in Equation (6.2), we defined a scenario as a vector $w = (\phi, t_{\text{rev}}, c_1, \ldots, c_n, T_1, \ldots, T_n, \eta_1, \ldots, \eta_m)$ where ϕ is the fuel scenario, t_{rev} is the time of the fuel scenario reveal, c_i is the development cost of technology i, T_i is the development time of technology i, and η_j is the CO_2 emission reduction performance of configuration j.

We defined the action set as $\mathcal{A} = \{\texttt{WAIT}, \texttt{D}_1, \ldots, \texttt{D}_n\}$ where `WAIT` is to do nothing in the current time step and $\texttt{D}_i$ is to start to develop technology i. Note that some actions may be invalid in some states. For state s, the set of valid actions $\mathcal{A}(s)$ was defined according to the following rules:

- `WAIT` is a valid action in all states.
- $\texttt{D}_i$ is a valid action in state s if the prerequisite technologies for technology i are already completed, and the development of technology i has not been started.

The prerequisite technologies are defined only for the upgrade of an ICE, namely `AtoMONO` and `HtoMONO`. These technologies cannot be developed unless the technologies from which each upgrade is made are completed. For instance, the agent cannot start developing `AtoMONO` unless `A&HO` is completed. We defined the development readiness function $f_{\text{preq},i}(\tau)\colon 2^{\{1,\ldots,n\}} \to \{\texttt{True}, \texttt{False}\}$ $(j = 1, \ldots, n)$ that represents whether the prerequisite technologies for technology i are completed given a set of completed technologies τ. The set of valid actions $\mathcal{A}(s)$ can be written as:

$$\mathcal{A}(s) \equiv \{\texttt{WAIT}\} \cup \{\texttt{D}_i \mid \ i = 1, \ldots, n;\ f_{\text{preq},i}(\tau(v))\} \tag{6.5}$$

where $\tau(v) \equiv \{i \mid \ v_i = \texttt{True}\}$ is the set of completed technologies.

6.3 Policy generation

6.3.1 Policy generation by reinforcement learning

Converting state, belief, and scenario into a vector

A tuple of a state, a belief, and a scenario (s, b, w) needs to be converted into a vector $\mathbf{x}$ to be fed to neural networks. State $s = (t, u_1, \ldots, u_n, v_1, \ldots, v_n)$ is converted into vector $\mathbf{x}_{\mathrm{s}}(s) = \begin{bmatrix} \mathbf{x}_{\mathrm{t}}(t)^\top & \mathbf{x}_{\mathrm{u}}(u)^\top & \mathbf{x}_{\mathrm{v}}(v)^\top \end{bmatrix}^\top$ where each vector is defined as:

$$\mathbf{x}_{\mathrm{t}}(t) = \begin{bmatrix} 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \end{bmatrix}^\top \tag{6.6a}$$

$$\mathbf{x}_{\mathrm{u}}(u) = \begin{bmatrix} u_1 & \ldots & u_n \end{bmatrix}^\top \tag{6.6b}$$

$$\mathbf{x}_{\mathrm{v}}(v) = \begin{bmatrix} v_1 & \ldots & v_n \end{bmatrix}^\top \tag{6.6c}$$

$\mathbf{x}_{\mathrm{t}}(t)$ is a vector whose elements are all zero except for the t-th element, which is one.

Belief b is a subset of $\mathcal{W}$, and its vectorization is not straightforward. Let us state that a belief b is an *ellipsoid belief* if and only if b can be expressed using vectors $\{\mathbf{x}_i\}_{i=1}^{d_{\mathrm{w}}}$ as:

$$b = \left\{ (w_1, \ldots, w_{d_{\mathrm{w}}}) \;\middle|\; \sum_{i}^{d_{\mathrm{w}}} (\beta_i(\mathbf{x}_i, w_i))^2 \leq 1 \right\} \tag{6.7}$$

where $\beta_i \colon \mathbb{R}^{d_{\mathrm{x},i}} \times \mathcal{W}_i \to \mathbb{R}$ is the budget-of-uncertainty function representing the budget of uncertainty that w_i takes given the parameters vector $\mathbf{x}_i \in \mathbb{R}^{d_{\mathrm{x},i}}$. $d_{\mathrm{x},i}$ is the dimension of vector $\mathbf{x}_i$:

$$\forall i \in \{1, \ldots, d_{\mathrm{w}}\} \colon d_{\mathrm{x},i} = 2 \tag{6.8}$$

One can confirm that the uncertainty model $\mathcal{U}(h)$ is an ellipsoid belief by defining β_i and $\mathbf{x}_i$ as

$$\beta_1(\mathbf{x}_1, w_1) \equiv \begin{cases} x_{11} & (\text{if } w_1 = \texttt{CRM}) \\ x_{12} & (\text{if } w_1 = \texttt{HA}) \end{cases} \tag{6.9a}$$

$$2 \leq i \leq d_{\mathrm{w}} \colon \beta_i(\mathbf{x}_i, w_i) \equiv \frac{2w_i - x_{i1} - x_{i2}}{x_{i2} - x_{i1}} \tag{6.9b}$$

$$\mathbf{x}_1 \equiv \begin{bmatrix} \frac{d_{\texttt{CRM}}}{h} & \frac{d_{\texttt{HA}}}{h} \end{bmatrix}^\top \tag{6.10a}$$

$$2 \leq i \leq d_{\mathrm{w}} \colon \mathbf{x}_i \equiv \begin{bmatrix} \tilde{w}_i - h\Delta w_i & \tilde{w}_i + h\Delta w_i \end{bmatrix}^\top \tag{6.10b}$$

where x_{ij} denotes the j-th element of vector $\mathbf{x}_i$. Let us assume we update an ellipsoid belief b expressed with $\{\mathbf{x}_i\}_{i=1}^{d_{\mathrm{w}}}$ by "fixing" w_k, and obtain b'. Formally

$$b' = \{w = (w_1, \ldots, w_{d_{\mathrm{w}}}) \mid \ w \in b, w_k = \bar{w}_k\} \tag{6.11}$$

Then the obtained belief b' is also an ellipsoid belief expressed with the budget-of-uncertainty function defined in Equation (6.9) and the parameters vector $\{\mathbf{x}'_i\}_{i=1}^{d_{\mathrm{w}}}$

$$\mathbf{x}'_1 = \begin{cases} \begin{bmatrix} 0 & \infty \end{bmatrix}^\top & (\text{if } k = 1 \wedge \bar{w}_k = \texttt{CRM}) \\ \begin{bmatrix} \infty & 0 \end{bmatrix}^\top & (\text{if } k = 1 \wedge \bar{w}_k = \texttt{HA}) \\ \dfrac{\mathbf{x}_1}{\bar{\beta}} & (\text{otherwise}) \end{cases} \tag{6.12a}$$

$$2 \leq i \leq d_{\mathrm{w}} \colon \mathbf{x}'_i = \begin{cases} \begin{bmatrix} \bar{w}_k & \bar{w}_k \end{bmatrix}^\top & (\text{if } k = i) \\ \begin{bmatrix} \dfrac{x_{i1} + x_{i2}}{2} - \bar{\beta}\dfrac{x_{i2} - x_{i1}}{2} & \dfrac{x_{i1} + x_{i2}}{2} + \bar{\beta}\dfrac{x_{i2} - x_{i1}}{2} \end{bmatrix}^\top & (\text{otherwise}) \end{cases} \tag{6.12b}$$

where

$$\bar{\beta} \equiv 1 - |\beta_k(\mathbf{x}_k, \bar{w}_k)| \tag{6.12c}$$

is the scaling factor of the belief ellipsoid. Finally, belief b is converted into a vector:

$$\mathbf{x}_{\mathrm{b}}(b) = \begin{bmatrix} \mathbf{x}_1^\top & \dots & \mathbf{x}_{d_{\mathrm{w}}}^\top \end{bmatrix}^\top \tag{6.13}$$

Note that elements in $\mathbf{x}_1$ are clipped into $[0, 1]$.

Scenario $w = (\phi, t_{\mathrm{rev}}, c_1, \dots, c_n, T_1, \dots, T_n, \eta_1, \dots, \eta_m)$ is converted into vector $\mathbf{x}_{\mathrm{w}}(w) = \begin{bmatrix} \mathbf{x}_{\upphi}^\top(\phi) & \mathbf{x}_{\mathrm{t_{rev}}}^\top(t_{\mathrm{rev}}) & \mathbf{x}_{\mathrm{c}}^\top(c) & \mathbf{x}_{\mathrm{T}}^\top(T) & \mathbf{x}_{\upeta}^\top(\eta) \end{bmatrix}^\top$ where each vector is defined as:

$$\mathbf{x}_{\upphi}(\phi) = \begin{cases} \begin{bmatrix} 1 & 0 \end{bmatrix}^\top & (\text{if } \phi = \texttt{CRM}) \\ \begin{bmatrix} 0 & 1 \end{bmatrix}^\top & (\text{otherwise}) \end{cases} \tag{6.14a}$$

$$\mathbf{x}_{\mathrm{t_{rev}}}(t_{\mathrm{rev}}) = \begin{bmatrix} t_{\mathrm{rev}} \end{bmatrix} \tag{6.14b}$$

$$\mathbf{x}_{\mathrm{c}}(c) = \begin{bmatrix} c_1 & \dots & c_n \end{bmatrix}^\top \tag{6.14c}$$

$$\mathbf{x}_{\mathrm{T}}(T) = \begin{bmatrix} T_1 & \dots & T_n \end{bmatrix}^\top \tag{6.14d}$$

$$\mathbf{x}_{\upeta}(\eta) = \begin{bmatrix} \eta_1 & \dots & \eta_m \end{bmatrix}^\top \tag{6.14e}$$

Finally, the vector that is fed to the neural network is

$$\mathbf{x}(s, b, w) = \begin{bmatrix} \mathbf{x}_{\mathrm{s}}^\top(s) & \mathbf{x}_{\mathrm{b}}^\top(b) & \mathbf{x}_{\mathrm{w}}^\top(w) \end{bmatrix}^\top \tag{6.15}$$

Learning the optimal policy with the deep reinforcement learning

We applied the deep Q-network algorithm to solve the maximin optimal Bellman equation. Table 6.4 lists hyperparameters, their search spaces, and their adopted values.

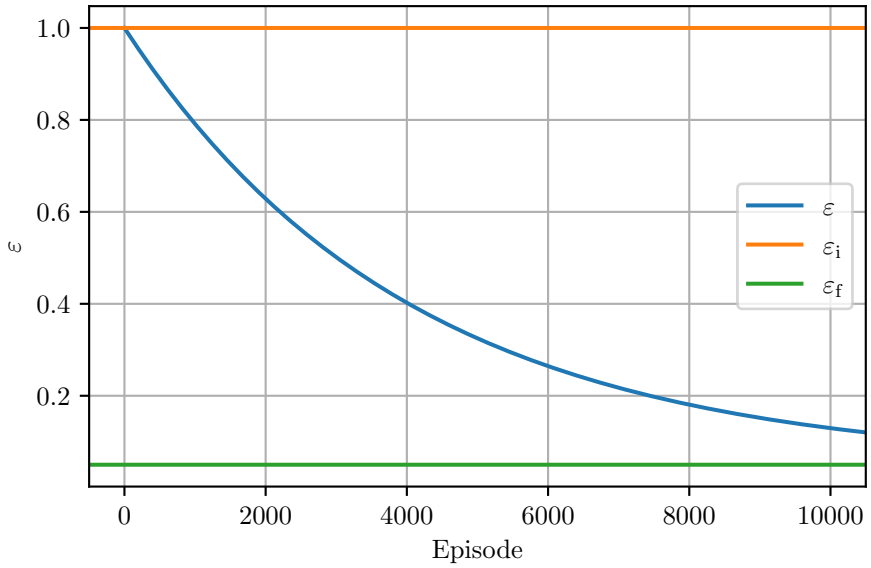


Figure 6.1 The scheduling of ε (blue). The orange and green lines show the initial and final value of ε, respectively. The horizontal axis shows the number of episodes.

The *neural network architecture* has two options: the vanilla Q-network (Figure 3.1) and the dueling network (Figure 3.2). If the *separate input* is `True`, only the connections within the state, belief, and scenario are connected in the initial n_{SL} fully connected layers to capture their features first before combining them.

If *scenario sampling in every step* is `True`, the scenario is sampled from belief b_t at every time step, otherwise only at the beginning of each episode. The *Pareto scenario sampling probability* defines the probability of sampling from the Pareto scenarios, and the *vertex probability* defines the probability of sampling from vertex scenarios if *select scenario from vertices* is `True`.

We normalized the reward by $\hat{r}_t \equiv \dfrac{r_t}{\bar{R}}$. The policy model used in training is the ε-greedy with the scheduling of ε. In the n_{ep}-th episode, the value of ε is set to:

$$\varepsilon = \varepsilon_{\text{f}} + (\varepsilon_{\text{i}} - \varepsilon_{\text{f}}) \exp\left(-\frac{n_{\text{ep}}}{n_{\varepsilon}}\right) \tag{6.16}$$

The histories of the worst cumulative reward under the obtained policy in the scenarios sampled from $\mathcal{U}(0)$ and $\mathcal{U}(1)$ are shown in Figures 6.2 and 6.3.

Table 6.4 Hyperparameters used in training.

Hyperparameter	Adopted value	Search space
Neural network parameters		
Neural network architecture	Dueling DQN	{DQN, Dueling DQN}
Separate input	`False`	{`True`, `False`}
The number of separate layers n_{SL}	—	$[1, 3]$
The number of hidden layers	4	$[2, 4]$
Activation function	ReLU	—
Hidden layer size	157	$[128, 256]$
Scenario sampling parameters		
Scenario sampling in every step	`False`	{`True`, `False`}
The number of scenario samples	26	$[20, 200]$
Pareto scenario sampling probability	0.56	$[0.5, 0.9]$
Select scenario from vertices	`True`	{`True`, `False`}
Vertex probability	0.83	$[0.1, 0.9]$
Training parameters		
Discount factor γ	0.99	$\{0.99, 1\}$
Reward scaling factor $\bar{R}$	91.4	$[50, 200]$
Initial value of ε: ε_{i}	1.0	—
Final value of ε: ε_{f}	0.05	—
Decay factor of ε in episodes : n_{ε}	4031.7	$[2500, 5000]$
Batch size	396	$[32, 128]$
Replay memory size	1332	$[1000, 10000]$
Target network update frequency	74	$[20, 200]$
Loss function	Huber loss [103]	—
Huber loss L1–L2 threshold	1	—
Optimizer	Adam [104]	—
Learning rate	7.61×10^{-3}	$[10^{-4}, 10^{-2}]$
Adam parameter (β_1, β_2)	(0.9, 0.999)	—
Adam parameter ϵ	10^{-8}	—
Test parameters		
The number of scenarios in the test	27	—

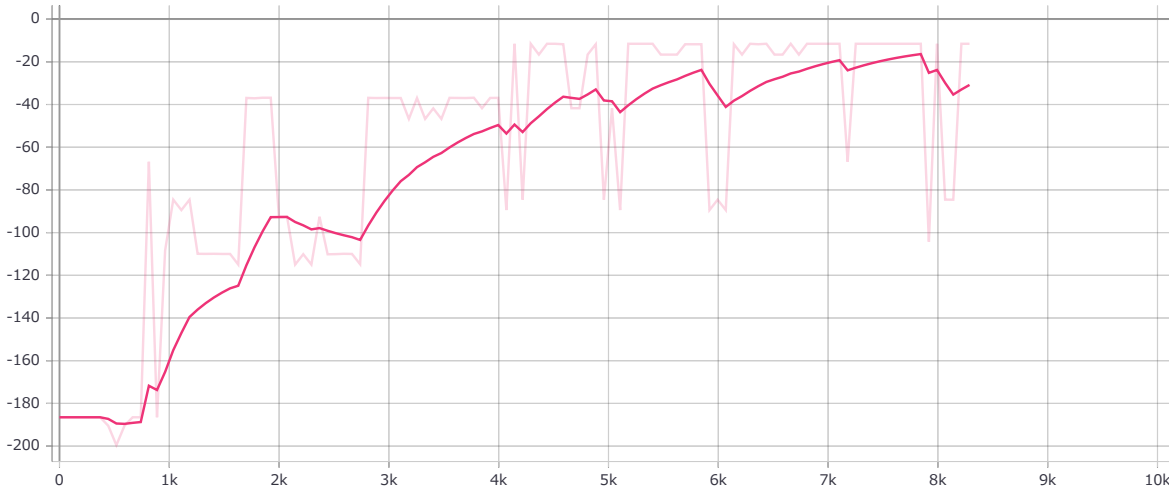


Figure 6.2 History of the cumulative reward in the test episode ($h = 0$). The horizontal axis shows the number of episodes.

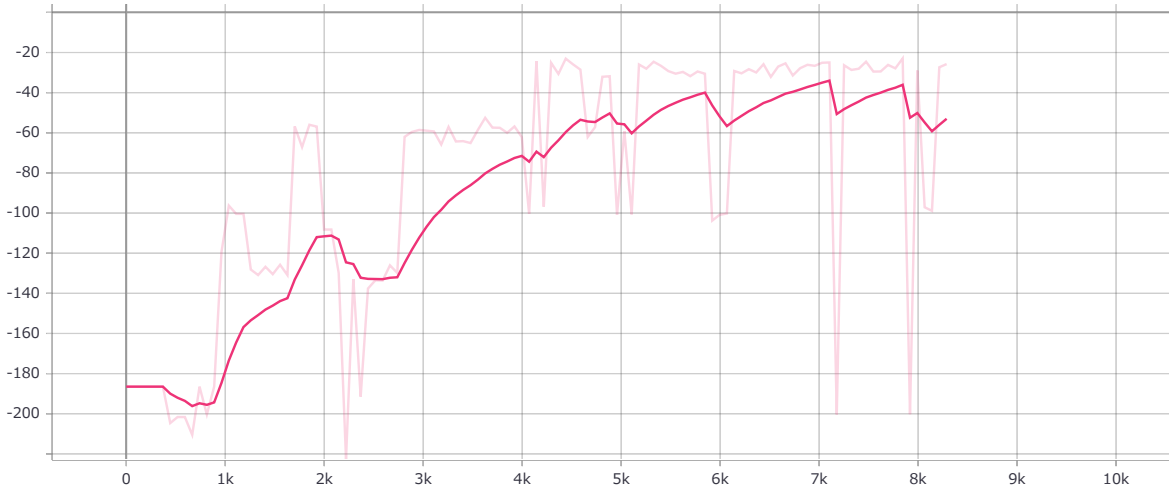


Figure 6.3 History of the cumulative reward in the test episode ($h = 1$). The horizontal axis shows the number of episodes.

Table 6.5 Parameters used in the HoU analysis of the `MarinePropulsion` problem.

Parameter	Value
Discount rate γ	1
Scenarios sampling method	Uniform Pareto scenarios sampling and vertices sampling
The number of Pareto scenarios samples	983
The number of Pareto vertices samples	17 ($= d_{\mathrm{w}}$)
The samples of h	25 evenly spaced samples in $[0, 1]$

6.3.2 Policy generation by experts

We defined five expert policies:

Ammonia Mixed The agent starts to develop `A&HO`.

Ammonia Mono The agent starts to develop `A&HO`. When `A&HO` is completed, it starts to develop `AtoMONO`.

Hydrogen Mixed The agent starts to develop `H&M`.

Hydrogen Mono The agent starts to develop `H&M`. When `H&M` is completed, it starts to develop `HtoMONO`.

CO_2 Collection The agent starts to develop `CC`.

6.4 HoU analysis and policy/HoU selection

6.4.1 HoU analysis settings

The parameters used in the HoU analysis are shown in Table 6.5. 25 values of the horizon of uncertainty h were sampled with even spaces in $[0, 1]$ ($h = 0, \frac{1}{24}, \ldots, \frac{23}{24}, 1$). Under each horizon of uncertainty h, we sampled 1,000 scenarios from $\mathcal{U}(h)$, 17 of which were the Pareto vertices of the ellipsoid-like region $\mathcal{U}(h)$, and the others were uniformly sampled from the Pareto-front (uniform Pareto scenarios sampling). In total, 25,000 scenarios were prepared.

The sampling of the Pareto vertices and the Pareto scenarios was conducted almost the same as in the other case study. See Section 5.4.1 for the details. The difference from the other case study is that the direction of goodness of the fuel scenario ϕ is not apparent. Therefore, one of the two fuel scenarios is randomly sampled as the worst or best value.

In addition to the five expert policies, we added a policy under which the agent always takes `WAIT` action. We simulated the six explicitly-defined policies under the 25,000 scenarios, both for the maximum and minimum cumulative reward, and the maximin RL policy for the minimum cumulative reward, resulting in 325,000 simulations in total.

6.4.2 HoU analysis results and discussion

The HoU plots with different performance measures are shown in Figures 6.4a to 6.4d. Findings from the HoU analysis are:

F.1 **CO_2 Collection** scores a higher cumulative reward than the other policies in the best-case scenarios in $h \geq 0.25$.

F.2 **Ammonia Mono** scores a higher cumulative reward than the other expert policies in the nominal and the worst-case scenarios.

F.3 The maximin RL policy scores the highest cumulative reward in the worst-case scenarios in any $h \in [0, 1]$.

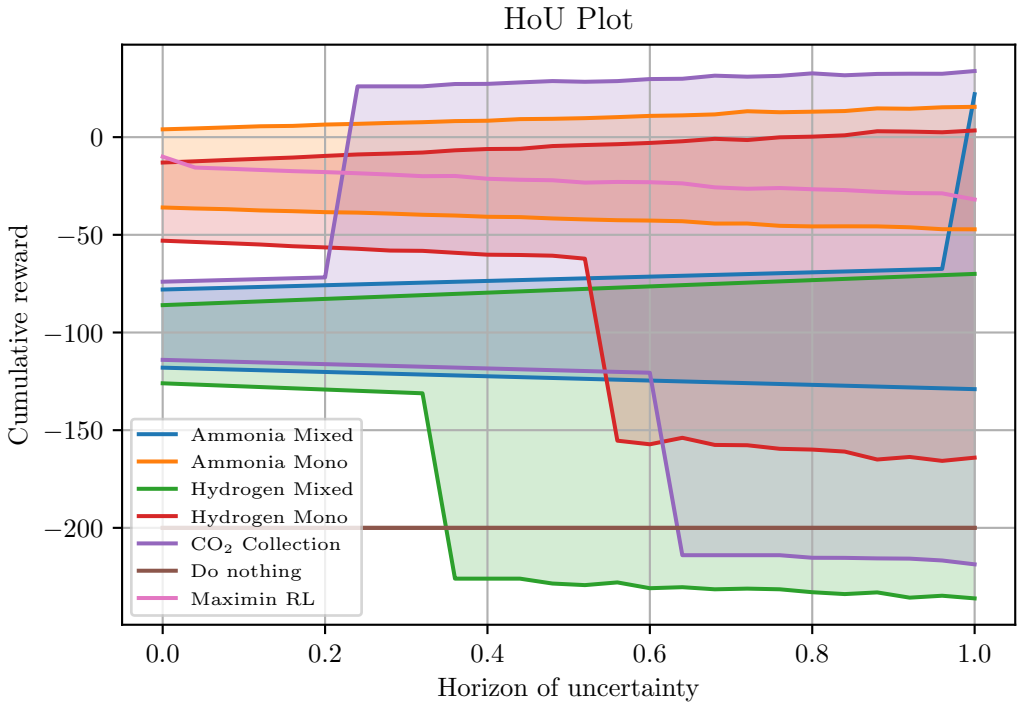
F.4 Only **Ammonia Mono**, **Hydrogen Mono**, and the maximin RL policy achieve both IMO targets in the nominal scenario ($h = 0$).

F.5 **Ammonia Mixed**, **Hydrogen Mixed**, **Hydrogen Mono**, and **CO_2 Collection** show "drop" in the worst-case cumulative reward.

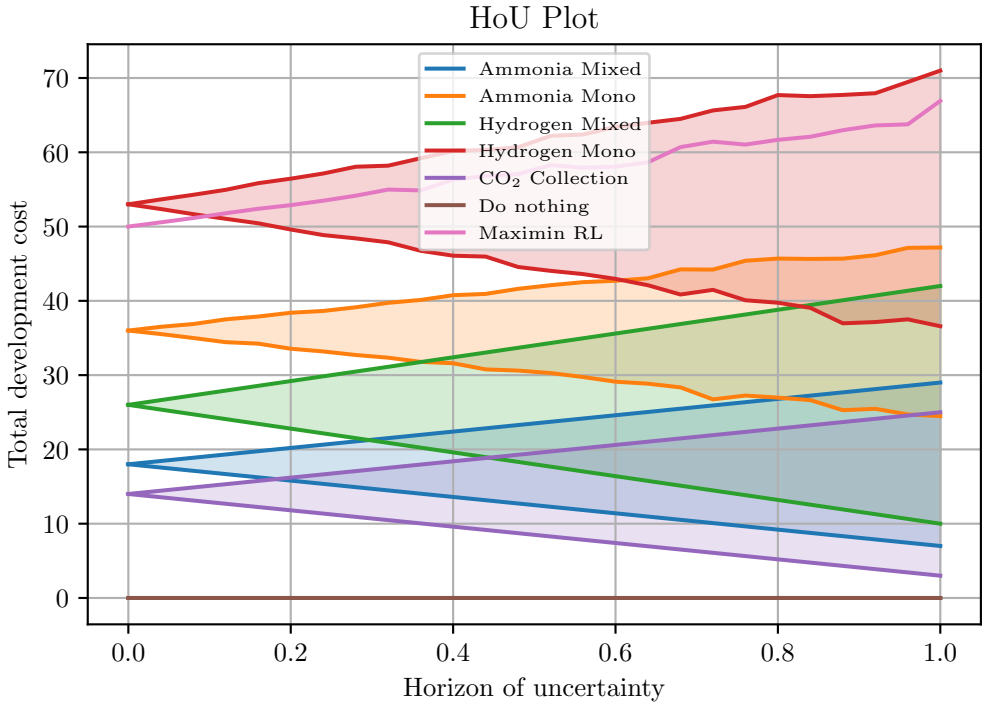
F.6 The range of the cumulative reward under each policy gradually increases as h increases except for the "drops."

F.7 Only **Ammonia Mono** and the maximin RL policy achieve both IMO targets in any $h \in [0, 1]$.

The reason for **F.1** is that the CO_2 collection technology has low development cost both at the lower and the upper bounds and has a possibility of achieving both IMO targets since the upper bound of CO_2 reduction performance of the CO_2 collection technology (0.95) well exceeds the IMO targets. However, as **F.2** suggests, in the nominal and the worst scenarios, **Ammonia Mono** performs the other expert policies. This is because in the nominal scenario, only **Ammonia Mono** and **Hydrogen Mono** achieve both IMO targets, and the development of configuration `H` costs more than that of configuration `A`. As observed in **F.3**, the reinforcement learning found a policy that outperforms the predefined expert policies. At any value of h, the agent following the maximin RL policy started developing technology `CC`, then technology `A&HO`, and technology `AtoMONO` after `A&HO` was completed. The maximin RL policy outperforms **Ammonia Mono** because it receives the majority bonus regardless of the fuel scenario. Also, it outperforms **CO_2 Collection** because it can achieve both IMO targets even if the CO_2 reduction performance $\eta_{\texttt{CC}}$ is not sufficient, and receives the majority bonus regardless of the fuel scenario. **F.4** is not surprising because the nominal CO_2 reduction performance of configurations `CC`, `A&HO`, and `H&M` are all below the IMO 2050 target of 70 %, while that of configurations `A` and `H` are both above the target. **F.5** shows that under the expert policies other than **Ammonia Mono**, either of the IMO targets becomes not guaranteed as h increases. In particular, under **Hydrogen Mono**, both IMO targets are achieved only when $h < 0.6$. The gradual increase in the performance range mentioned in **F.6** is due to the uncertainty in each technology's development cost. As discussed in Section 5.4.2, the gradual increase is linear in the horizon of uncertainty. The reason for **F.7** is that configuration `A` is the only configuration that can achieve both IMO targets. Although the CO_2 reduction performance of `H` ($\eta_{\texttt{H}}$) exceeds both IMO targets even at the lower bound, not both targets are achieved for reasons discussed later in the scenario analysis.

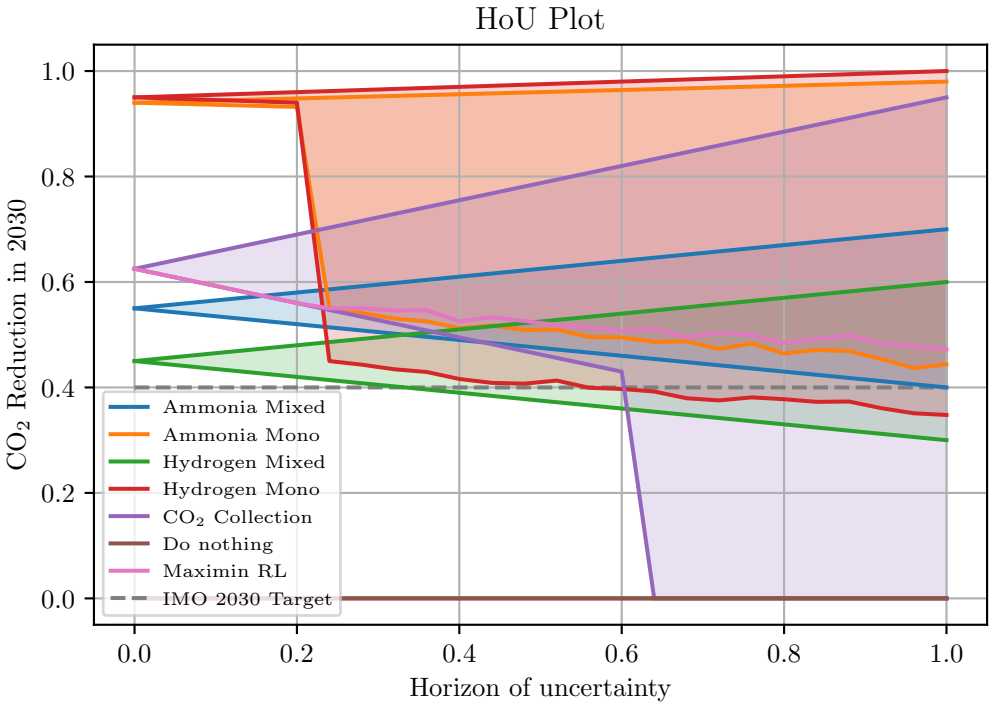


(a) Cumulative reward

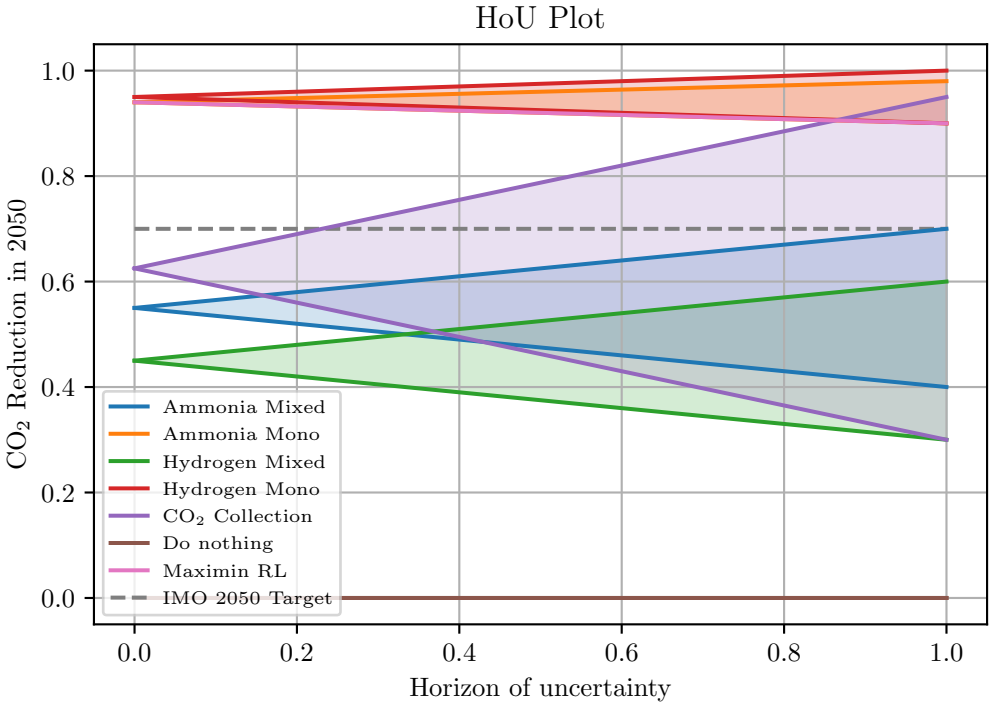


(b) Total technology development cost

Figure 6.4 The HoU plot of the expert policies and the maximin RL policy. Only the worst-case performance measure is plotted for the maximin RL policy.



(c) CO_2 reduction in 2030.



(d) CO_2 reduction in 2050.

Figure 6.4 The HoU plot of the expert policies and the maximin RL policy. Only the worst-case performance measure is plotted for the maximin RL policy (cont.).

Table 6.6 Parameters used in the scenario analysis of the `MarinePropulsion` problem.

Parameter	Value
Discount rate γ	1
Scenarios sampling method	Uniform scenarios sampling
The number of scenarios	10,000

6.5 Scenario analysis

6.5.1 Scenario analysis settings

The parameters used in the scenario analysis are shown in Table 6.6. 10,000 scenarios were sampled quasi-uniformly from $\mathcal{U}(1)$. Under each scenario, the six candidate policies were simulated, resulting in 60,000 simulations in total. The performance measures for each simulation were: the cumulative reward and whether both IMO targets were achieved.

The scenarios were sampled from $\mathcal{U}(1)$ as follows. Let n_{w} the number of scenarios to sample. First, sample uniformly distributed n_{w} random points $\{\mathbf{x}^{(i)}\}_{i=1}^{n_{\mathrm{w}}}$ inside a d_{w}-dimensional unit ball by the steps described in Section 5.5.1. For each point $\mathbf{x}^{(i)}$, x_{ij} is allocated to the j-th uncertain parameter as its "budget of uncertainty." For $j = 1$, i.e., the dimension corresponds to the fuel scenario, w_1 is randomly sampled from set $\{\phi \in \{\texttt{CRM}, \texttt{HA}\} \mid \ d_\phi \leq |x_{i1}|\}$. For $j \neq 1$, w_j is sampled in the same way as the other case study. If $x_{ij} = 0$, then the uncertain parameter will be set to its nominal value. Otherwise, there are two w_j's that satisfy $d_j(w_j) = |x_{ij}|$. Let us denote the two solutions as $w_j^{(1)}$ and $w_j^{(2)}$, and assume without the loss of generality that $w_j^{(1)}$ is better than $w_j^{(2)}$. For example, $w_j^{(1)} < w_j^{(2)}$ if w_j is a cost. Then the uncertain parameter will be set to the better value if $x_{ij} > 0$, or to the worse value if $x_{ij} < 0$.

6.5.2 Scenario analysis results and discussion

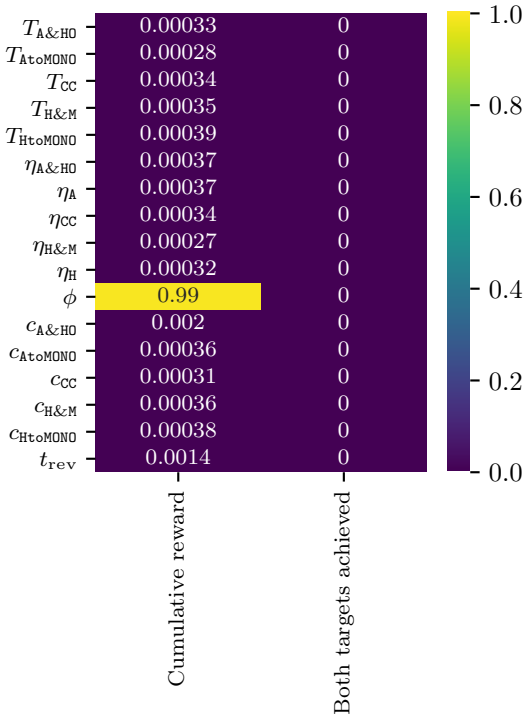
We analyzed each uncertain parameter's sensitivity against each performance measure by calculating the feature score in the regression using the extremely randomized trees

[79]. The scores under each policy are shown in Figure 6.5. It can be observed that under **Ammonia Mixed**, **Ammonia Mono**, **Hydrogen Mixed**, and the maximin RL policy, every score against the target achievement is zero. This is because whether both IMO targets are achieved is the same across all the scenarios. It can also be seen that under such policies, the fuel scenario ϕ has the dominant sensitivity against the cumulative reward, except for the maximin RL policy. The reason for the dominant sensitivity is that the majority bonus is designed to be larger than the deviation in the development cost. Therefore, the results may be different with a different reward function design. The maximin RL policy is not affected as much by the fuel scenario because the agent develops both configurations `A` and `CC`, and receives the majority bonus regardless of the fuel scenario.

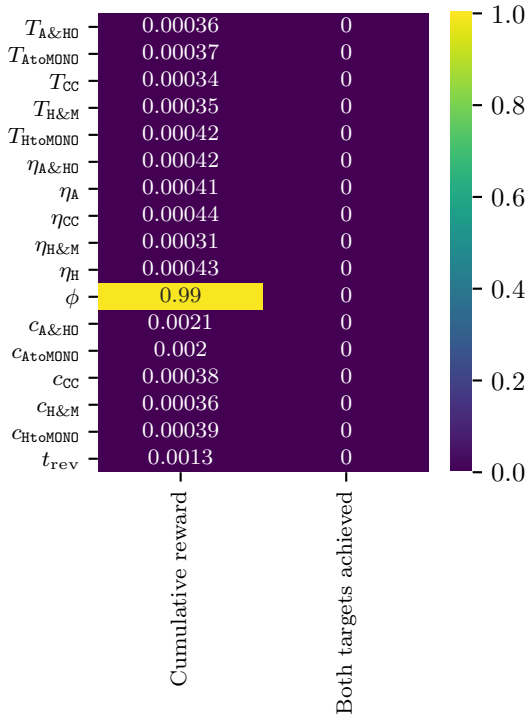
We applied the scenario discovery to see when both IMO targets are achieved and when not under **Hydrogen Mono** and **CO_2 Collection**. We defined the cases of interest (CoI) as the simulation cases where both IMO targets were achieved. The Pareto front in the density–coverage space obtained with the PRIM under each policy is shown in Figure 6.6. Under **Hydrogen Mono**, the box with the maximum density and the one with the maximum coverage are close to each other, indicating the cases of interest and the other cases are separated in the scenario space.

Figure 6.7 shows the distribution of the cases of interest and the other cases in the restricted dimension of the box with the largest density. Under **Hydrogen Mono**, configuration `H&M`'s CO_2 reduction performance is the restricted dimension, indicating that the risk under **Hydrogen Mono** is that the mixed hydrogen ICE's CO_2 reduction performance may miss the IMO 2030 target. Under **CO_2 Collection**, there are two restricted dimensions: $T_{\texttt{CC}}$ and $\eta_{\texttt{CC}}$. The box shows that if the development time of technology `CC` is shorter than some threshold and configuration `CC`'s CO_2 reduction performance is better than some threshold, then **CO_2 Collection** can achieve both IMO targets.

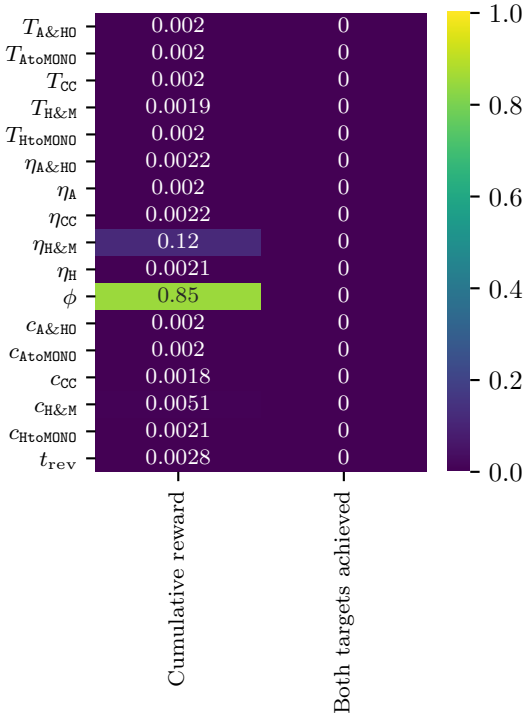
Figure 6.8 shows the regional sensitivity to observe each uncertain parameter's sensitivity against whether both IMO targets were achieved. One can observe that



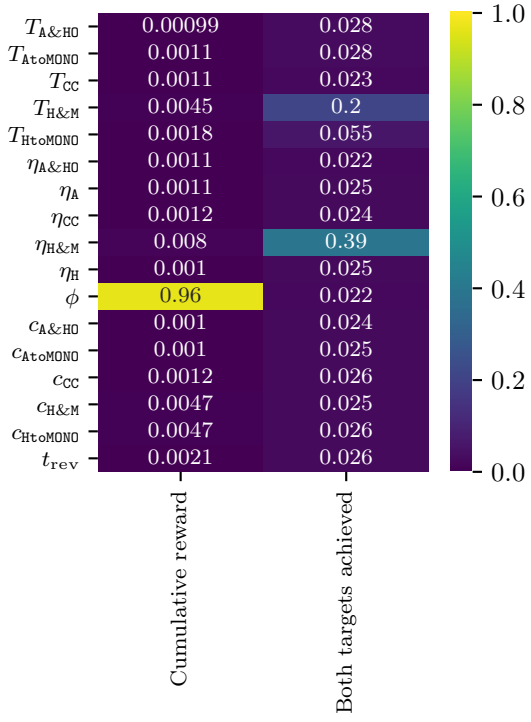
(a) Policy: Ammonia Mixed



(b) Policy: Ammonia Mono

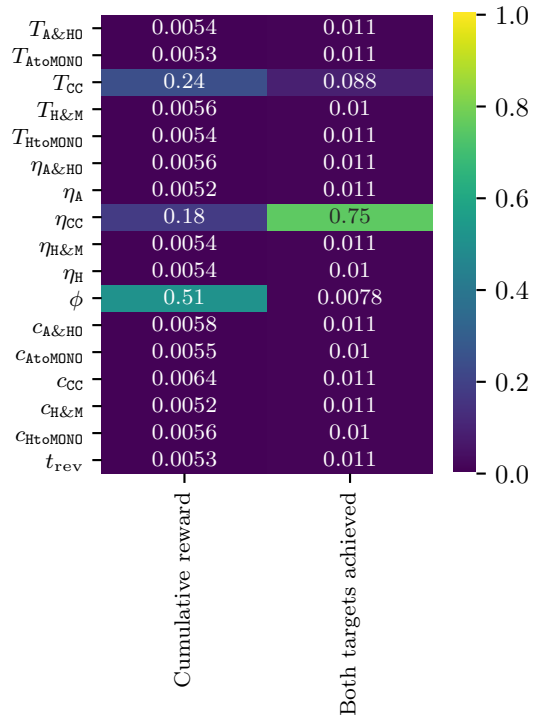


(c) Policy: Hydrogen Mixed

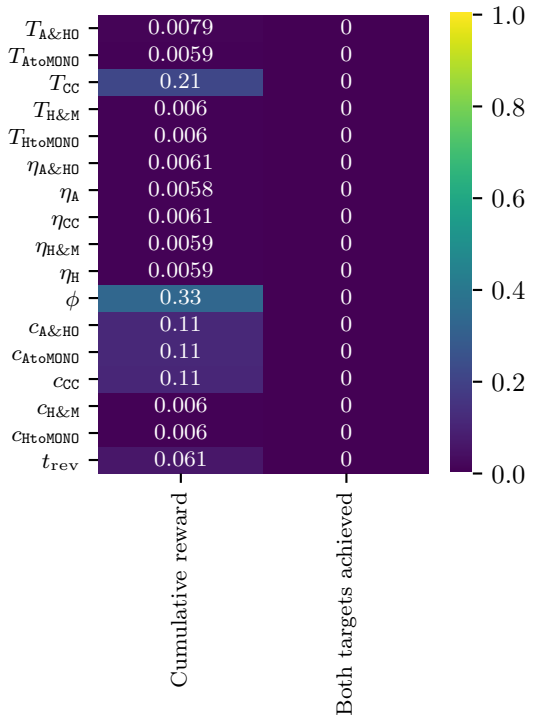


(d) Policy: Hydrogen Mono

Figure 6.5 Feature scoring of each uncertain parameter under each policy.

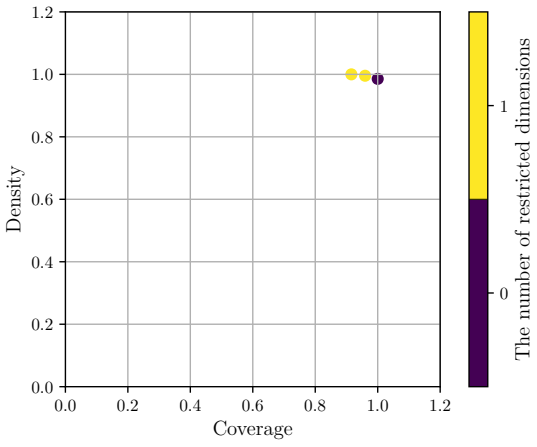


(e) Policy: CO_2 Collection

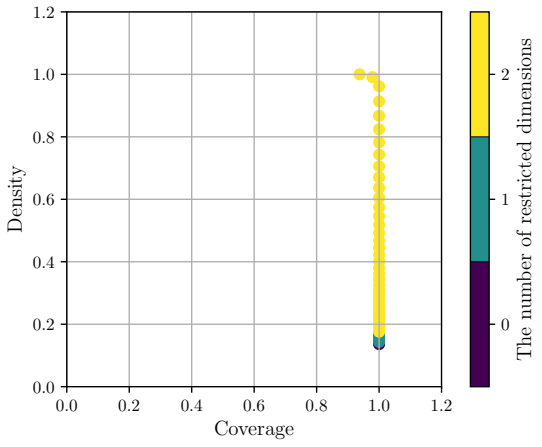


(f) Policy: CO_2 Collection and Ammonia Mono (Maximin RL policy)

Figure 6.5 Feature scoring of each uncertain parameter under each policy (cont.).



(a) Policy: Hydrogen Mono



(b) Policy: CO_2 Collection

Figure 6.6 Density–coverage Pareto front of each policy. The cases of interest were defined as the ones where all of the three missions are conductible at the final time step.

under **Hydrogen Mono**, in addition to $\eta_{\texttt{H\&M}}$, whose high sensitivity was also mentioned in the feature scoring analysis, $T_{\texttt{H\&M}}$ and $T_{\texttt{HtoMONO}}$ also have sensitivity. This is because if the development time of technology `H&M` and that of `HtoMONO` are short enough for configuration `H` to be available before 2030, the IMO 2030 target will be achieved because the CO_2 reduction performance of configuration `H` is 0.9 in the worst case. The uncertain parameters that have high sensitivity under **CO_2 Collection** are, as also shown in the feature scoring analysis, $T_{\texttt{CC}}$ and $\eta_{\texttt{CC}}$.

6.6 Discussion

Table 6.7 summarizes each policy's advantages and disadvantages based on the findings from the HoU analysis and the scenario analysis. The final decision is up to the decision-maker's attitude toward each performance measure and risk.

If the decision-maker is interested in achieving both IMO targets inexpensively, even if the chance of target achievement is compromised, **CO_2 Collection** may be the choice because its best-case cumulative reward is better than that of other policies in $h \geq 0.25$.

If the decision-maker is interested in achieving both IMO targets even at $h = 1$, but allows the risk of missing the majority bonus when $\phi = \texttt{CRM}$, then **Ammonia Mono** may be the choice.

Suppose the decision-maker is interested in achieving both IMO targets even at $h = 1$ and does not risk missing the majority bonus. In that case, the maximin RL policy may be the choice because the decision-maker can achieve both IMO targets and receive the majority bonus in any scenario $w \in \mathcal{U}(1)$.

Under the current assumption of uncertainty, the other policies may not be the choice. The policies involving the hydrogen ICE performed worse than that of the ammonia ICE largely because the uncertainty in the CO_2 reduction performance of configuration `H&M` is so large that it can not achieve the IMO 2030 target.

Table 6.7 Advantages and disadvantages of each policy.

Policy	Advantages	Disadvantages
Ammonia Mixed	• The IMO 2030 target is achieved in all cases.	• The IMO 2050 target is *not* achieved in all cases except for the best scenario at $h = 1$. • There is a risk of $\phi = \texttt{CRM}$.
Ammonia Mono	• Both IMO targets are achieved in all cases.	• There is a risk of $\phi = \texttt{CRM}$.
Hydrogen Mixed	• The IMO 2030 target is achieved under small h.	• The IMO 2050 target is *not* achieved in all cases. • The IMO 2030 target is *not* achieved under large h due to its poor performance. • There is a risk of $\phi = \texttt{CRM}$.
Hydrogen Mono	• Both IMO targets are achieved under small h.	• The IMO 2030 target is *not* achieved under large h due to the poor performance of Hydrogen Mixed. • There is a risk of $\phi = \texttt{CRM}$.
CO_2 Collection	• Both IMO targets may be achieved under large h.	• Neither IMO targets may be achieved under large h due to its poor performance. • There is a risk of $\phi = \texttt{HA}$.
CO_2 Collection & Ammonia Mono (Maximin RL policy)	• Both IMO targets are achieved in all cases. • The decision-maker can prepare for both fuel scenarios.	• One of the developed configurations is *retrospectively* unnecessary.

6.7 Expert feedback

We shared the results with experts working in the maritime industry and obtained the following feedback:

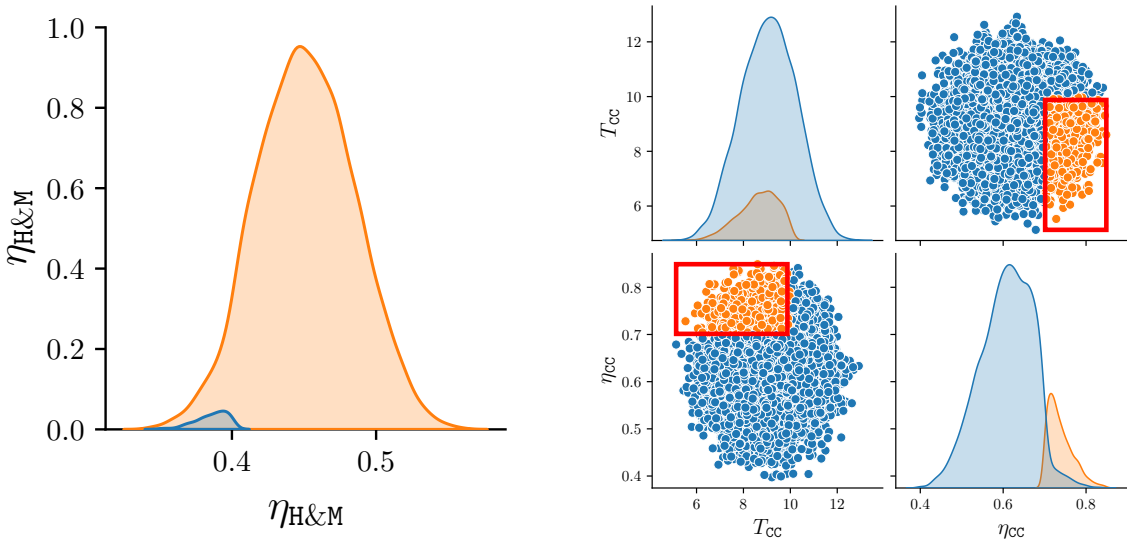
Benefits from the framework

- The HoU plot is useful to consider the problem from various stakeholders' perspectives, such as start-up companies that are willing to take risks to gain a large reward and policymakers that want to minimize the risk of not achieving the targets.
- The conclusion that hydrogen-based strategies have large uncertainty does not contradict their mental model.
- It helps focus the decision-makers' attention to dominant strategies, namely **CO_2 Collection** and **Ammonia Mono** in the `MarinePropulsion` problem. It is usually difficult to eliminate strategy options based only on a qualitative discussion.

Opportunities for enhancement

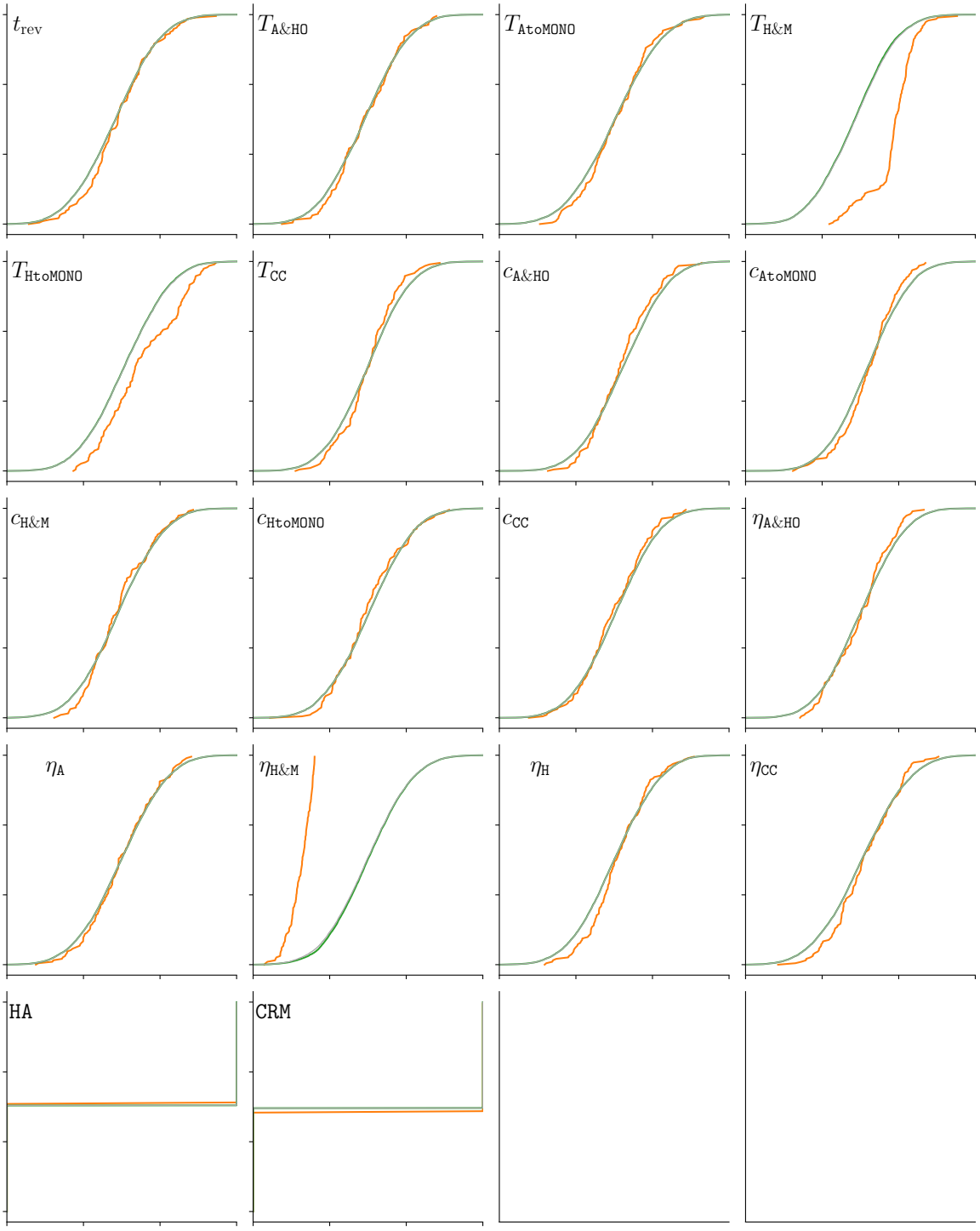
- The actual decision-making is not only about the discrete options, but rather about the balance between the discrete options, such as the fraction of investment in each technology development. This is not modeled in the current model.
- The actual decision-makers, especially ones in private companies, consider how their competitors will act in the decision-making process. This is not modeled in the current framework.

It can be safely said that the results from the MSRDM framework can benefit an actual decision-making problem, though there still exist some possible enhancements to be made, especially in the expressiveness of the model.



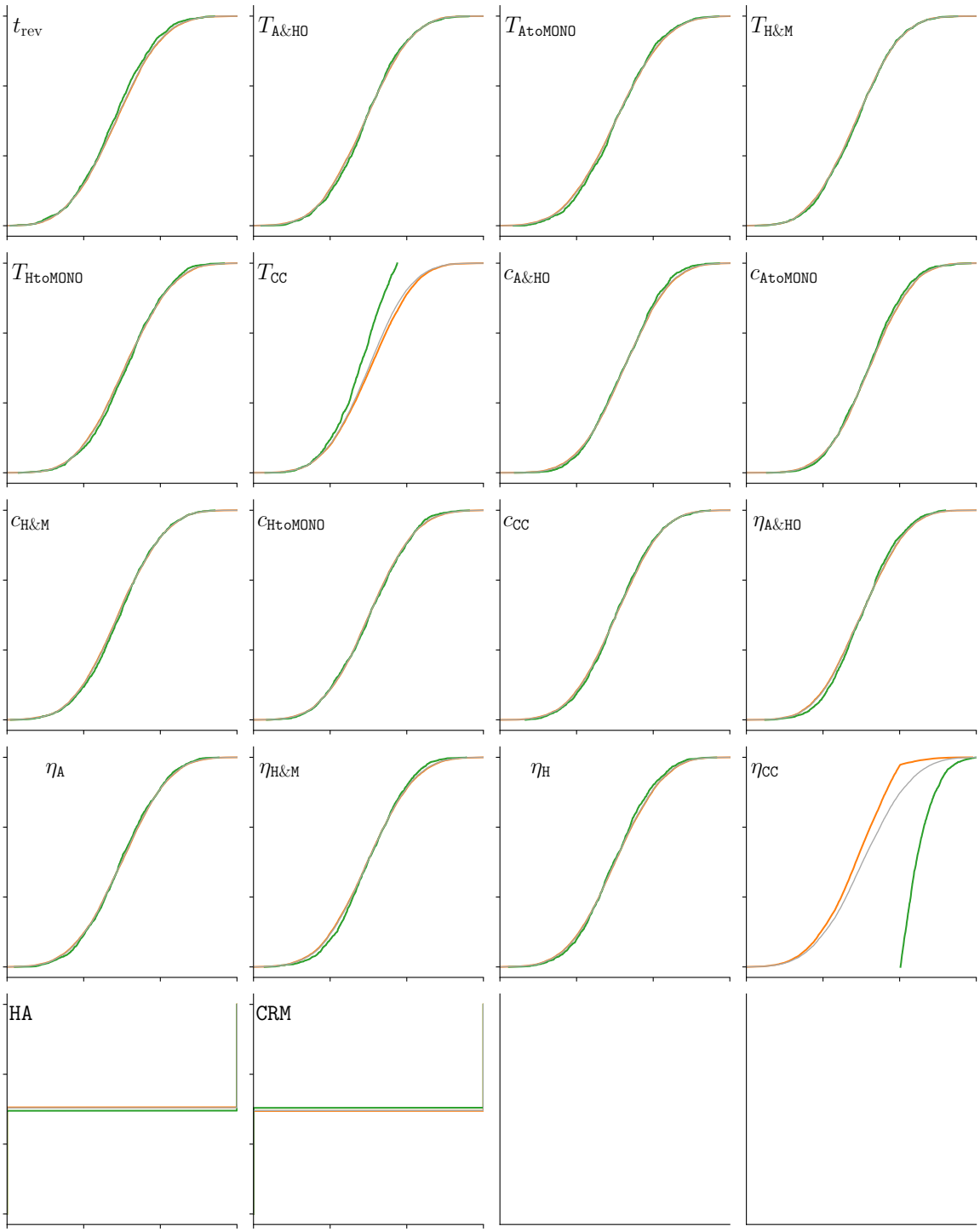
(a) Policy: Hydrogen Mono (b) Policy: CO_2 Collection

Figure 6.7 Pair plots in the restricted dimension in the scenario discovery of each policy.



(a) Policy: Hydrogen Mono

Figure 6.8 The regional sensitivity analysis of the cases of interest under each policy.



(b) Policy: CO_2 Collection

Figure 6.8 The regional sensitivity analysis of the cases of interest under each policy (cont.).

Chapter 7

Conclusions

In this dissertation, we proposed three concepts:

- a multi-stage-robust-decision-making Markov decision process (MSRDM-MDP)
- the horizon-of-uncertainty (HoU) analysis
- the multi-stage robust decision making (MSRDM) framework

Chapter 3 presented the definition of an MSRDM-MDP and proved that the maximax and maximin policies can be obtained by solving the maximax and maximin Bellman equations, respectively, using a reinforcement learning algorithm such as the deep Q-network (DQN).

Chapter 4 presented the MSRDM framework and the HoU analysis as an analysis method used in the framework. The detailed process was described using a toy problem, `SimpleMining`, and showed that the framework was able to help understand the trade-off between the robustness and the performance of each policy option.

We then applied the proposed framework to the two case studies. Chapter 5 demonstrated the framework in the technology roadmapping of the space formation flying system, and analyzed policy options with different development staging strategies. Chapter 6 demonstrated the technology roadmapping of the marine propulsion system, and analyzed strategies to reduce maritime CO_2 emission. In both case studies,

the analyses' results made the advantages and disadvantages of each policy option clear. It was also confirmed by feedback from experts that the proposed framework is capable of benefiting an actual decision-making problem, despite some opportunities for enhancement.

As validated by the case studies, the proposed MSRDM framework has three functions: to model deep uncertainty, to provide robustness–performance trade-off, and to model multi-stage decision making. Existing decision-making frameworks have some functions, but none has all the functions to the best of our knowledge, as shown in Table 1.2.

However, note that our framework still has limitations that may be addressed in the future. Notable limitations include:

- Our framework cannot support cases where the uncertainty is so large that even the uncertainty model cannot be constructed.
- Our framework cannot provide stochastic guarantees about the outcome because the uncertainty is treated as non-probabilistic.
- The maximax and maximin Bellman equations assume a single performance measure—the discounted cumulative reward—and cannot consider multiple performance measures and trade-off among them.
- Our framework cannot model game theoretic environments where multiple agents act to maximize their performance measures.

Some functions were not verified in this dissertation, but are likely to be handled with existing methods and algorithms. Such functions include:

- To calculate the best/worst scenario more efficiently using optimization algorithms in the reinforcement learning process or the HoU analysis.
- To solve the maximax and maximin Bellman equations for MSRDM-MDPs with continuous or hybrid actions.

- To model an environment with not only non-probabilistic uncertainty but also probabilistic uncertainty.

References

[1] V. Pop, *Who Owns the Moon? Extraterrestrial Aspects of Land and Mineral Resources Ownership*. 2009.

[2] P. De Man, *Exclusive Use in an Inclusive Environment - The Meaning of the Non-Appropriation Princilple for Space Resource Exploitation*. 2016.

[3] E. W. Paxson, III, "Sharing the benefits of outer space ExplorSharing the benefits of outer space exploration: Space law and ation: Space law and economic DeEconomic development velopment," *Michigan Journal of International Law*, vol. 14, no. 3, pp. 487–517, 1993.

[4] G. Leterre, "Providing a legal framework for sustainable space mining activities," Master's thesis, Université du Luxembourg, Sept. 2017.

[5] E. Beauvois and G. Thirion, "Partial ownership for outer space resources," *Advances in Astronautics Science and Technology*, vol. 3, pp. 29–36, June 2020.

[6] Nishimura Institute of Advanced Legal Studies, the Space Resource Development Laws Study Group, "Nishimura institute of advanced legal studies report by the space resource development laws study group," tech. rep., Nishimura Institute of Advanced Legal Studies, Dec. 2016.

[7] D. Kornuta, A. Abbud-Madrid, J. Atkinson, J. Barr, G. Barnhard, D. Bienhoff, B. Blair, V. Clark, J. Cyrus, B. DeWitt, C. Dreyer, B. Finger, J. Goff, K. Ho, L. Kelsey, J. Keravala, B. Kutter, P. Metzger, L. Montgomery, P. Morrison, C. Neal, E. Otto, G. Roesler, J. Schier, B. Seifert, G. Sowers, P. Spudis, M. Sundahl, K. Zacny, and G. Zhu, "Commercial lunar propellant architecture: A collaborative study of lunar propellant production," tech. rep., 2018.

[8] S. Mallick and R. P. Rajagopalan, "If space is 'the province of mankind', who owns its resources?: An examination of the potential of space mining and its legal implications," tech. rep., Observer Research Foundation, Jan. 2019.

[9] J. F. Connolly, "Constellation program overview," Oct. 2006.

[10] J. L. Rattigan, D. J. Neubek, L. Dale Thomas, and C. Stegemoeller, "Constellation program lessons learned; volume i: Executive summary," Tech. Rep. NASA/SP-2011-6127-VOL-1, NASA Lyndon B. Johnson Space Center, 2011.

[11] National Aeronautics and Space Administration, "NASA's journey to mars, pioneering next steps in space exploration," Tech. Rep. NP-2015-08-2018-HQ, National Aeronautics and Space Administration, 2015.

[12] M. Gates, S. Stich, M. McDonald, B. Muirhead, D. Mazanek, P. Abell, and P. Lopez, "The asteroid redirect mission and sustainable human exploration," *Acta astronautica*, vol. 111, pp. 29–36, 2015.

[13] M. Gates, B. Muirhead, B. Naasz, M. McDonald, D. Mazanek, S. Stich, P. Chodas, and J. Reuter, "NASA's asteroid redirect mission concept development summary," in *2015 IEEE Aerospace Conference*, pp. 1–13, Mar. 2015.

[14] N. Strange, D. Landau, T. McElrath, G. Lantoine, and T. Lam, "Overview of mission design for NASA asteroid redirect robotic mission concept." Oct. 2013.

[15] C. Moore, "Technology development for NASA's asteroid redirect mission," *65th International Astronautical Congress, IAC-14-D2*, 2014.

[16] D. D. Mazanek, R. G. Merrill, J. R. Brophy, and R. P. Mueller, "Asteroid redirect mission concept: A bold approach for utilizing space resources," *Acta astronautica*, vol. 117, pp. 163–171, 2015.

[17] D. D. Mazanek, R. G. Merrill, S. P. Belbin, D. M. Reeves, K. D. Earle, B. J. Naasz, and P. A. Abell, "Asteroid redirect robotic mission : Robotic boulder capture option overview," no. August, pp. 1–22, 2014.

[18] National Aeronautics and Space Administration, "The artemis plan: NASA's lunar exploration program overview," Tech. Rep. NP-2020-05-2853-HQ, National Aeronautics and Space Administration, Sept. 2020.

[19] United States Government Accountability Office, "NASA assessments of major projects," Tech. Rep. GAO-20-405, United States Government Accountability Office, Apr. 2020.

[20] United States Government Accountability Office, "Space transportation: The content and uses of shuttle cost estimates," Tech. Rep. GAO/NSIAD-93-115, United States Government Accountability Office, Jan. 1993.

[21] T. Wu, "Shuttle programme lifetime cost," *Nature*, vol. 472, no. 7341, p. 38, 2011.

[22] J. M. Logsdon, "The space shuttle program: A policy failure?," *Science*, vol. 232, pp. 1099–1105, May 1986.

[23] S. Li, P. G. Lucey, R. E. Milliken, P. O. Hayne, E. Fisher, J.-P. Williams, D. M. Hurley, and R. C. Elphic, "Direct evidence of surface exposed water ice in the lunar polar regions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, pp. 8907–8912, Sept. 2018.

[24] "ispace." https://ispace-inc.com/. Accessed: 2021-1-4.

[25] M. Webster, S. Paltsev, J. Parsons, J. Reilly, and H. Jacoby, "Uncertainty in greenhouse gas emissions and costs of atmospheric stabilization," Tech. Rep. 165, The MIT Joint Program on the Science and Policy of Global Change, Nov. 2008.

[26] E. Marland, J. Cantrell, K. Kiser, G. Marland, and K. Shirley, "Valuing uncertainty part i: the impact of uncertainty in GHG accounting," *Carbon Management*, vol. 5, pp. 35–42, Feb. 2014.

[27] L. R. Kump, "Reducing uncertainty about carbon dioxide as a climate driver," *Nature*, vol. 419, pp. 188–190, Sept. 2002.

[28] J. Yuan, S. H. Ng, and W. S. Sou, "Uncertainty quantification of CO2 emission reduction for maritime shipping," *Energy policy*, vol. 88, pp. 113–130, Jan. 2016.

[29] P. R. Rose, "Dealing with risk and uncertainty in exploration: How can we improve?," *AAPG Bulletin*, vol. 71, pp. 1–16, Jan. 1987.

[30] M. Amram and N. Kulatilaka, "Uncertainty: the new rules for strategy," *The Journal of business strategy*, vol. 20, no. 3, pp. 25–29, 1999.

[31] I. Lerche, *Geological Risk and Uncertainty in Oil Exploration*. Academic Press, 1997.

[32] L. Benkherouf and J. A. Bather, "Oil exploration: Sequential decisions in the face of uncertainty," *Journal of applied probability*, vol. 25, no. 3, pp. 529–543, 1988.

[33] S. B. Suslick, D. Schiozer, and M. R. Rodriguez, "Uncertainty and risk analysis in petroleum exploration and production," *Terrae incognitae*, vol. 6, no. 1, p. 2009, 2009.

[34] T. Roach, Z. Kapelan, R. Ledbetter, and M. Ledbetter, "Comparison of robust optimization and Info-Gap methods for water resource management under deep uncertainty," *Journal of Water Resources Planning and Management*, vol. 142, Sept. 2016.

[35] N. K. Ajami, G. M. Hornberger, and D. L. Sunding, "Sustainable water resource management under hydrological uncertainty," *Water resources research*, vol. 44, Nov. 2008.

[36] W. E. Walker, M. Haasnoot, and J. H. Kwakkel, "Adapt or perish: A review of planning approaches for adaptation under deep uncertainty," *Sustainability (Switzerland)*, vol. 5, no. 3, pp. 955–979, 2013.

[37] G. H. Huang and D. P. Loucks, "An inexact two stage stochastic programming model for water resources management under uncertainty," *Civil Engineering and Environmental Systems*, vol. 17, no. 2, pp. 95–118, 2000.

[38] K. W. Hipel and Y. Ben-Haim, "Decision making in an uncertain world: information-gap modeling in water resources management," *IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews: a publication of the IEEE Systems, Man, and Cybernetics Society*, 1999.

[39] P. R. Taylor, "The mismeasure of risk," in *Handbook of Risk Theory: Epistemology, Decision Theory, Ethics, and Social Implications of Risk* (S. Roeser, ed.), pp. 441–475, Springer Science & Business Media, 2012.

[40] N. N. Taleb, *The black swan: The impact of the highly improbable*, vol. 2. Random house, 2007.

[41] F. H. Knight, *Risk, uncertainty and profit*, vol. 31. Houghton Mifflin, 1921.

[42] S. F. LeRoy and L. D. Singell, "Knight on risk and uncertainty," *The journal of political economy*, vol. 95, pp. 394–406, Apr. 1987.

[43] R. J. Lempert, S. W. Popper, and S. C. Bankes, *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis*. 2003.

[44] W. E. Walker and P. J. T. M. Bloemen, *Decision Making under Deep Uncertainty*. 2019.

[45] Y. Ben-Haim, "Innovation, optimization and their dilemmas: An Info-Gap perspective." Coping with Uncertainty: Normative Approaches, Current Practice, 2017.

[46] D. J. Kessler, N. L. Johnson, J. C. Liou, and M. Matney, "The kessler syndrome: implications to future space operations," *Advances in the Astronautical Sciences*, vol. 137, no. 8, p. 2010, 2010.

[47] S. Savage, "The flaw of averages," *Harvard Business Review*, Nov. 2002.

[48] S. L. Savage and H. M. Markowitz, *The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty*. John Wiley & Sons, June 2009.

[49] The Economist, "Cutting the cord." https://www.economist.com/special-report/1999/10/07/cutting-the-cord, Oct. 1999. Accessed: 2020-12-31.

[50] L. Whitaker, "Ads unplugged," *American demographics*, vol. 23, no. 6, pp. 30–33, 2001.

[51] A. Saran, K. Cruthirds, and M. S. Minor, "Ad acceptance: Scale development, purification, and validation of cell telephone advertising acceptance," in *Revolution in Marketing: Market Driving Changes*, pp. 62–66, Springer International Publishing, 2015.

[52] E. Crawley, B. Cameron, and D. Selva, *System Architecture: Strategy and Product Development for Complex Systems*. USA: Prentice Hall Press, 1st ed., 2015.

[53] S. C.-H. Yang, D. M. Wolpert, and M. Lengyel, "Theoretical perspectives on active sensing," *Current opinion in behavioral sciences*, vol. 11, pp. 100–108, Oct. 2018.

[54] R. de Neufville and S. Scholtes, *Flexibility in Engineering Design*. MIT Press, Aug. 2011.

[55] A. Guma, J. Pearson, K. Wittels, R. de Neufville, and D. Geltner, "Vertical phasing as a corporate real estate strategy and development option," *Journal of Corporate Real Estate*, vol. 11, pp. 144–157, Jan. 2009.

[56] Goettsch Partners, "300 east randolph." https://www.gpchicago.com/architecture/300-east-randolph/. Accessed: 2021-1-2.

[57] J. R. Pearson and K. S. Wittels, *Real options in action : vertical phasing in commercial real estate development*. PhD thesis, Massachusetts Institute of Technology, 2008.

[58] A. C. a. C. Guma, *A real options analysis of a vertically expandable real estate development*. PhD thesis, Massachusetts Institute of Technology, 2008.

[59] O. L. de Weck, R. de Neufville, and M. Chaize, "Staged deployment of communications satellite constellations in low earth orbit," *Journal of Aerospace Computing, Information, and Communication*, vol. 1, pp. 119–136, Mar. 2004.

[60] R. J. Lempert, S. W. Popper, D. G. Groves, N. Kalra, J. R. Fischbach, S. C. Bankes, B. P. Bryant, M. T. Collins, K. Keller, and A. Hackbarth, "Making good decisions without predictions: Robust decision making for planning under deep uncertainty," tech. rep., RAND Corporation, 2013.

[61] Y. Ben-Haim, "Uncertainty, probability and information-gaps," *Reliability Engineering & System Safety*, 2004.

[62] W. B. Haskell and R. Jain, "A convex analytic approach to Risk-Aware markov decision processes," *SIAM Journal on Control and Optimization*, vol. 53, no. 3, pp. 1569–1598, 2015.

[63] R. Tyrrell Rockafellar and S. Uryasev, "Optimization of conditional Value-at-Risk," *Journal of risk*, vol. 2, pp. 21–42, Apr. 2000.

[64] Y. Yamai and T. Yoshiba, "Comparative analyses of expected shortfall and Value-at-Risk (2): Expected utility maximization and tail risk," *Monetary and Economic Studies*, vol. 20, pp. 95–115, Apr. 2002.

[65] D. Bertsimas and A. Thiele, "Robust and Data-Driven optimization: Modern decision making under uncertainty," *Models, Methods, and Applications for Innovative Decision Making*, no. March, pp. 95–122, 2006.

[66] D. Bertsimas, D. B. Brown, and C. Caramanis, "Theory and applications of robust optimization," *SIAM Review*, vol. 53, no. 3, pp. 464–501, 2011.

[67] R. A. Howard, *Dynamic programming and markov processes*. John Wiley, 1960.

[68] T. Morimura, *Reinforcement Learning*. Machine Learning Professional Series, Kodansha, Sept. 2019.

[69] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, May 1992.

[70] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

[71] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[72] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*, pp. 1995–2003, PMLR, 2016.

[73] F. C. Lunenburg, "The decision making process," *National Forum of Educational Administration And Supervision Journal*, vol. 27, no. 4, 2010.

[74] G. Forgionne and J. Newman, "An experiment on the effectiveness of creativity enhancing decision-making support systems," *Decision support systems*, vol. 42, pp. 2126–2136, Jan. 2007.

[75] J. J. Elam and M. Mead, "Can software influence creativity?," *Information Systems Research*, vol. 1, pp. 1–22, Mar. 1990.

[76] N. Althuizen and A. Reichel, "The effects of IT-Enabled cognitive stimulation tools on creative problem solving: A dual pathway to creativity," *Journal of Management Information Systems*, vol. 33, pp. 11–44, Jan. 2016.

[77] K. Wang and J. V. Nickerson, "A literature review on individual creativity support systems," *Computers in human behavior*, vol. 74, pp. 139–151, 2017.

[78] M. P. Van Der Burg and A. J. Tyre, "Integrating info-gap decision theory with robust population management: A case study using the mountain plover," *Ecological applications: a publication of the Ecological Society of America*, vol. 21, no. 1, pp. 303–312, 2011.

[79] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[80] J. H. Friedman and N. I. Fisher, "Bump hunting in High-Dimensional data," *Statistics and Computing*, vol. 9, no. 2, pp. 123–143, 1999.

[81] G. Krieger, A. Moreira, H. Fiedler, I. Hajnsek, M. Werner, M. Younis, and M. Zink, "TanDEM-X: A satellite formation for High-Resolution SAR interferometry," *IEEE transactions on geoscience and remote sensing: a publication of the IEEE Geoscience and Remote Sensing Society*, vol. 45, pp. 3317–3341, Nov. 2007.

[82] P. Bodin, R. Noteborn, R. Larsson, and J.-C. Berges, "Prisma formation flying demonstrator: Overview and conclusions from the nominal mission," *Advances in the Astronautical Sciences*, vol. 144, Feb. 2012.

[83] M. Kirschner, O. Montenbruck, and S. Bettadpur, "Flight dynamics aspects of the GRACE formation flight," pp. 1–8, 2001.

[84] "GRACE." https://grace.jpl.nasa.gov/mission/grace/. Accessed: 2021-1-6.

[85] DLR, "TerraSAR-X and TanDEM-X flying in close formation."

[86] K. Hayashida, K. Asakura, T. Yoneyama, T. Hanasaka, A. Ishikura, S. Sakuma, H. Noda, K. Okazaki, M. Hanoka, S. Ide, K. Hattori, H. Matsumoto, H. Tsunemi, H. Nakajima, H. Awaki, and J. S. Hiraga, "Formation flight mission of Multi-Image x-ray interferometer modules (MIXIM) for spatially resolved x-ray images around super massive black holes," in *Proceedings of 63rd Space Sciences and Technology Conference*, no. 1N06.

[87] A. N. Parmar, G. Hasinger, M. Arnaud, X. Barcons, D. Barret, A. Blanchard, H. Boehringer, M. Cappi, A. Comastri, T. Courvoisier, and others, "XEUS: the x-ray evolving universe spectroscopy mission," vol. 4851, pp. 304–313, 2003.

[88] T. Matsuo, S. Ikari, S. Ishiwata, H. Kondo, I. Kawano, and T. Ito, "Space infrared stellar interferometer for high spatial resolution of milli-arcsecond," in *Proceedings of 63rd Space Sciences and Technology Conference*, no. 1N05.

[89] S. Martin, D. P. Scharf, R. Wirz, O. Lay, D. McKinstry, B. Mennesson, G. Purcell, J. Rodriguez, L. Scherr, J. R. Smith, and L. Wayne, "Design study for a Planet-Finding space interferometer," in *2008 IEEE Aerospace Conference*, pp. 1–19, ieeexplore.ieee.org, Mar. 2008.

[90] M. Ando and B-DECIGO group, "Space Gravitational-Wave antenna: B-DECIGO," in *Proceedings of 63rd Space Sciences and Technology Conference*, no. 1N04, 2019.

[91] K. Danzmann and A. Rüdiger, "LISA technology—concept, status, prospects," *Classical and Quantum Gravity*, vol. 20, p. S1, Apr. 2003.

[92] T. Ito, S. Ikari, S. Sakai, and I. Kawano, "Strategic analysis on Japan's formation flying activities," in *Proceedings of 63rd Space Sciences and Technology Conference*, no. 1N02, 2019.

[93] United States Department of the Navy, United States Department of the Air Force, United States Marine Corps, United States Missile Defense Agency, and National Aeronautics and Space Administration, *Joint Agency Cost Schedule Risk and Uncertainty Handbook*. Mar. 2014.

[94] G. Marsaglia, "Choosing a point from the surface of a sphere," *Annals of Mathematical Statistics*, vol. 43, pp. 645–646, Apr. 1972.

[95] International Maritime Organization, "Third IMO GHG study 2014," tech. rep., International Maritime Organization, 2015.

[96] International Maritime Organization, "Adoption of the initial IMO strategy on reduction of GHG emissions from ships and existing IMO activity related to reducing GHG emissions in the shipping sector," 2018.

[97] The Japan Ship Technology Research Association and The Ministry of Land, Infrastructure, Transport and Tourism of Japan, "Roadmap to zero emission from international shipping," tech. rep., Shipping Zero Emission Project, Mar. 2020.

[98] N. d. Vries, *Safe and effective application of ammonia as a marine fuel*. PhD thesis, 2019.

[99] K. Kim, G. Roh, W. Kim, and K. Chun, "A preliminary study on an alternative ship propulsion system fueled by ammonia: Environmental and economic assessments," *Journal of marine science and engineering*, vol. 8, p. 183, Mar. 2020.

[100] M. M. El gohary, "Design of marine hydrogen internal combustion engine," *AEJ - Alexandria Engineering Journal*, vol. 48, pp. 57–65., Jan. 2009.

[101] I. S. Seddiek, M. M. Elgohary, and N. R. Ammar, "The hydrogen-fuelled internal combustion engines for marine applications with a case study," *Brodogradnja*, vol. 66, pp. 23–38, Mar. 2015.

[102] International Maritime Organization, "Update on IMO's work to address GHG emissions from fuel used for international shipping," Dec. 2019.

[103] P. J. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, pp. 73–101, Mar. 1964.

[104] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014.

Appendix A

Discussion on the Uncertainty Model $\mathcal{U}(h)$

A.1 Background

Consider an elliptic uncertainty model of uncertain parameters $(w_1, w_2) \in \mathbb{R}^2$ and the update of belief b_t to b_{t+1} by observing w_1, shown in Figure 4.4. Observing w_1 will result in change in the lower and upper bounds of w_2. Formally,

$$\inf \{w_2 \mid \exists w_1 \in \mathbb{R} \colon (w_1, w_2) \in b_{t+1}\} \geq \inf \{w_2 \mid \exists w_1 \in \mathbb{R} \colon (w_1, w_2) \in b_t\} \tag{A.1a}$$

$$\sup \{w_2 \mid \exists w_1 \in \mathbb{R} \colon (w_1, w_2) \in b_{t+1}\} \leq \sup \{w_2 \mid \exists w_1 \in \mathbb{R} \colon (w_1, w_2) \in b_t\} \tag{A.1b}$$

In terms of probability, observing a random variable does not affect the knowledge on another independent random variable. Therefore it seems reasonable to consider a rectangular uncertainty model:

$$\mathcal{U}(h) \equiv \left[w_1^{\mathrm{LB}}, w_1^{\mathrm{UB}}\right] \times \left[w_2^{\mathrm{LB}}, w_2^{\mathrm{UB}}\right] \tag{A.2}$$

However, one can confirm that the value of the prospecting action is zero under the rectangular uncertainty model because the optimal action is $\mathtt{M}_{i^*}$ where

$$i^* \equiv \underset{i \in \{1,2\}}{\operatorname{argmax}}\, w_i^{\mathrm{LB}} \tag{A.3}$$

In this appendix, we will analyze the value of prospecting actions assuming some probability distribution of uncertain parameters, and show that assuming ellipsoid uncertainty models can approximate the prospecting action value in cases where the agents maximizes the α-percentile of the reward.

Note that the following discussion only considers two-stage problems where the agent first prospects some value and chooses one option based on the observation. Applying the same discussion to multi-stage problems can be difficult because the optimal policy to maximize α-percentile risk measure needs to depend on not only the current state but also the history [62].

A.2 Problem assumptions

We consider a single-stage problem with a discrete decision variable $i \in \{1, \ldots, n\}$. We assume that the reward that the agent receives after selecting i, denoted as $W_i \in \mathbb{R}$, is a random variable with the probability density function $f_i(W_i)$, and variables $\{W_i\}_{i=1}^n$ are mutually independent. We denote the cumulative distribution function of W_i as $F_i(W_i)$.

Let R be the random variable representing the reward. We assume that the agent's risk measure is α-percentile R_α:

$$R_\alpha = \sup \{r \in \mathbb{R} \mid \Pr(R < r) < \alpha\} \tag{A.4}$$

Without additional information, the optimal decision i^* that maximizes R_α is the one that has the maximum α-percentile:

$$i^* = \operatorname*{argmax}_{i\in\{1,\dots,n\}} F_i^{-1}(\alpha) \tag{A.5a}$$

$$R_\alpha^* = \max_{i\in\{1,\dots,n\}} F_i^{-1}(\alpha) \tag{A.5b}$$

A.3 Value of a prospecting action

We now consider a prospecting action. Without loss of generality, we assume that the agent knows the value of W_1. After W_1 is becomes known, what will be the optimal decision i^* that maximizes the α-percentile?

Let $\bar{w}_1$ the value of W_1 that becomes known. It is clearly the optimal policy to select $i = 1$ if $\bar{w}_1$ is larger than some threshold ρ and $i \neq 1$ otherwise. Because the choice of $i \neq 1$ in case $\bar{w}_1 \leq \rho$ may depend on $\bar{w}_1$, let us define $\hat{q}(\bar{w}_1)\colon \mathbb{R} \to \{2, 3, \dots, n\}$ representing the decision the agent selects if $\bar{w} \leq \rho$. Then the reward is

$$R = \begin{cases} W_1 & (\text{if } W_1 > \rho) \\ W_{\hat{q}(W_1)} & (\text{otherwise}) \end{cases} \tag{A.6}$$

Let us calculate the cumulative distribution function of R. When $r \leq \rho$, the probability density function of R is calculated as:

$$\begin{aligned} &\frac{\Pr(r \leq R \leq r + \mathrm{d}r)}{\mathrm{d}r} \\ &= \frac{1}{\mathrm{d}r} \sum_{i\in\{2,3,\dots,n\}} \Pr(W_1 \leq \rho) \Pr(\hat{q}(W_1) = i \mid W_1 \leq \rho) \Pr(r \leq W_i \leq r + \mathrm{d}r) \\ &= \sum_{i\in\{2,3,\dots,n\}} F_1(\rho) Q_i(\rho) f_i(r) \end{aligned} \tag{A.7}$$

where

$$Q_i(\rho) \equiv \Pr(\hat{q}(W_1) = i \mid W_1 \leq \rho) \in [0, 1] \tag{A.8}$$

Note that $\forall\rho\colon\ \sum_{i=2}^{n} Q_i(\rho) = 1$. Its derivative, denoted as $q_i(\rho)$, can be calculated as:

$$\begin{aligned}
q_i(\rho) &\equiv \lim_{\Delta\rho\to 0} \frac{Q_i(\rho+\Delta\rho) - Q_i(\rho)}{\Delta\rho} \\
&= \lim_{\Delta\rho\to 0} \frac{\Pr(\hat{q}(W_1) = i \mid W_1 \le \rho+\Delta\rho) - \Pr(\hat{q}(W_1) = i \mid W_1 \le \rho)}{\Delta\rho} \\
&= \lim_{\Delta\rho\to 0} \frac{1}{\Delta\rho}\Big(\Pr(\hat{q}(W_1) = i \mid W_1 \le \rho) \Pr(W_1 \le \rho \mid W_1 \le \rho+\Delta\rho) \\
&\quad + \Pr(\hat{q}(W_1) = i \mid \rho \le W_1 \le \rho+\Delta\rho) \Pr(\rho \le W_1 \le \rho+\Delta\rho \mid W_1 \le \rho+\Delta\rho) \\
&\quad - \Pr(\hat{q}(W_1) = i \mid W_1 \le \rho)\Big) \\
&= \lim_{\Delta\rho\to 0} \frac{1}{\Delta\rho}\Big(\Pr(\hat{q}(W_1) = i \mid W_1 \le \rho) \frac{F_1(\rho)}{F_1(\rho+\Delta\rho)} \\
&\quad + \Pr(\hat{q}(W_1) = i \mid \rho \le W_1 \le \rho+\Delta\rho) \frac{F_1(\rho+\Delta\rho) - F_1(\rho)}{F_1(\rho+\Delta\rho)} \\
&\quad - \Pr(\hat{q}(W_1) = i \mid W_1 \le \rho)\Big) \\
&= \frac{f_1(\rho)}{F_1(\rho)} \left([\hat{q}(\rho) = i] - Q_i(\rho)\right)
\end{aligned} \tag{A.9}$$

Note that $[\cdot]$ is the Iverson bracket.

Then the cumulative distribution function is its integral:

$$\begin{aligned}
\Pr(R \le r) &= \int_{-\infty}^{r} \Pr(r' \le R \le r' + \mathrm{d}r') \\
&= \int_{-\infty}^{r} \sum_{i\in\{2,3,\ldots,n\}} F_1(\rho) Q_i(\rho) f_i(r') \mathrm{d}r' \\
&= \sum_{i\in\{2,3,\ldots,n\}} F_1(\rho) Q_i(\rho) F_i(r)
\end{aligned} \tag{A.10}$$

When $\rho < r$,

$$\begin{aligned}
&\Pr(r \le R \le r + \mathrm{d}r) \\
&= \Pr(r \le W_1 \le r + \mathrm{d}r) \\
&\quad + \sum_{i \in \{2,3,\ldots,n\}} \Pr(W_1 \le \rho) \Pr(\hat{q}(W_1) = i \mid W_1 \le \rho) \Pr(r \le W_i \le r + \mathrm{d}r) \\
&= f_1(r)\mathrm{d}r + \sum_{i \in \{2,3,\ldots,n\}} F_1(\rho) Q_i(\rho) f_i(r)\mathrm{d}r
\end{aligned} \tag{A.11}$$

Then the cumulative distribution function is its integral:

$$\begin{aligned}
\Pr(R \le r) &= \int_{-\infty}^{r} \Pr(r' \le R \le r' + \mathrm{d}r') \\
&= \int_{-\infty}^{\rho} \sum_{i \in \{2,3,\ldots,n\}} F_1(\rho) Q_i(\rho) f_i(r')\mathrm{d}r' \\
&\quad + \int_{\rho}^{r} \left(f_1(r') + \sum_{i \in \{2,3,\ldots,n\}} F_1(\rho) Q_i(\rho) f_i(r') \right) \mathrm{d}r' \\
&= F_1(r) - F_1(\rho) + \sum_{i \in \{2,3,\ldots,n\}} F_1(\rho) Q_i(\rho) F_i(r)
\end{aligned} \tag{A.12}$$

Equations (A.10) and (A.12) give the cumulative distribution function of R:

$$F_R(r \mid \rho, \{Q_i\}_{i=2}^n) = \begin{cases} \displaystyle\sum_{i \in \{2,3,\ldots,n\}} F_1(\rho) Q_i(\rho) F_i(r) & (\text{if } r \le \rho) \\ \displaystyle F_1(r) - F_1(\rho) + \sum_{i \in \{2,3,\ldots,n\}} F_1(\rho) Q_i(\rho) F_i(r) & (\text{otherwise}) \end{cases} \tag{A.13}$$

Note that it is continuous at $r = \rho$. The limit at $r = \rho$ can be calculated as:

$$\alpha_\rho(\rho) \equiv \lim_{r \to \rho} F_R(r \mid \rho, \{Q_i\}_{i=2}^n) = \sum_{i \in \{2,3,\ldots,n\}} F_1(\rho) Q_i(\rho) F_i(\rho) \tag{A.14}$$

$\alpha_\rho(\rho)$ is monotonically increasing:

$$\frac{\mathrm{d}\alpha_\rho}{\mathrm{d}\rho} = \sum_{i \in \{2,3,\ldots,n\}} (f_1(\rho)[\hat{q}(\rho) = i] F_i(\rho) + F_1(\rho) Q_i(\rho) f_i(\rho)) \ge 0 \tag{A.15}$$

Because $\lim_{\rho\to-\infty}\alpha_\rho(\rho)=0$ and $\lim_{\rho\to+\infty}\alpha_\rho(\rho)=1$, $\exists\hat{\rho}\in\mathbb{R}$ s.t. $\alpha_\rho(\hat{\rho})=\alpha$.

The decision variables that determine R's cumulative distribution function and α-percentile are ρ and $\{Q_i\}_{i=2}^n$. Consider optimizing ρ under fixed $\{Q_i\}_{i=2}^n$ to maximize R's α-percentile R_α.

When $\rho\geq\hat{\rho}$, $\alpha_\rho(\rho)\geq\alpha_\rho(\hat{\rho})=\alpha$. Therefore, $R_\alpha\leq\rho$, and the following equation holds:

$$\sum_{i\in\{2,3,\ldots,n\}}F_1(\rho)Q_i(\rho)F_i(R_\alpha)=\alpha \tag{A.16}$$

The derivatives of both sides of Equation (A.16) with respect to ρ are:

$$\sum_{i\in\{2,3,\ldots,n\}}\left(f_1(\rho)Q_i(\rho)F_i(R_\alpha)+F_1(\rho)q_i(\rho)F_i(R_\alpha)+F_1(\rho)Q_i(\rho)f_i(R_\alpha)\frac{\mathrm{d}R_\alpha}{\mathrm{d}\rho}\right)=0 \tag{A.17}$$

Equation (A.9) can be substituted into Equation (A.17), which can be rearranged to

$$\begin{aligned}\frac{\mathrm{d}R_\alpha}{\mathrm{d}\rho}&=-\frac{f_1(\rho)\sum_{i\in\{2,3,\ldots,n\}}[\hat{q}(\rho)=i]F_i(R_\alpha)}{F_1(\rho)\sum_{i\in\{2,3,\ldots,n\}}Q_i(\rho)f_i(R_\alpha)}\\&\leq 0\end{aligned} \tag{A.18}$$

Therefore, R_α decreases as ρ increases from $\rho=\hat{\rho}$.

When $\rho\leq\hat{\rho}$, $\alpha_\rho(\rho)\leq\alpha_\rho(\hat{\rho})=\alpha$. Therefore, $R_\alpha\geq\rho$, and the following equation holds:

$$F_1(R_\alpha)-F_1(\rho)+\sum_{i\in\{2,3,\ldots,n\}}F_1(\rho)Q_i(\rho)F_i(R_\alpha)=\alpha \tag{A.19}$$

The derivatives of both sides of Equation (A.19) with respect to ρ are:

$$\begin{aligned}&f_1(R_\alpha)\frac{\mathrm{d}R_\alpha}{\mathrm{d}\rho}-f_1(\rho)\\&+\sum_{i\in\{2,3,\ldots,n\}}\left(f_1(\rho)Q_i(\rho)F_i(R_\alpha)+F_1(\rho)q_i(\rho)F_i(R_\alpha)+F_1(\rho)Q_i(\rho)f_i(R_\alpha)\frac{\mathrm{d}R_\alpha}{\mathrm{d}\rho}\right)=0\end{aligned} \tag{A.20}$$

Equation (A.9) can be substituted into Equation (A.20), which can be rearranged to

$$\begin{aligned}\frac{\mathrm{d}R_\alpha}{\mathrm{d}\rho} &= \frac{f_1(\rho)\left(1-\sum_{i\in\{2,3,\ldots,n\}}[\hat{q}(\rho)=i]F_i(R_\alpha)\right)}{f_1(R_\alpha)+\sum_{i\in\{2,3,\ldots,n\}}F_1(\rho)Q_i(\rho)f_i(R_\alpha)} \\ &\geq 0\end{aligned} \tag{A.21}$$

Therefore, R_α decreases as ρ decreases from $\rho=\hat{\rho}$, and $\hat{\rho}$ is the optimal threshold to maximize the α-percentile R_α under fixed $\{Q_i\}_{i=2}^n$. Let $\hat{R}_\alpha$ be the α-percentile under $\rho=\hat{\rho}$. Then the following equation holds:

$$\hat{R}_\alpha = \hat{\rho} \tag{A.22a}$$

$$\sum_{i\in\{2,3,\ldots,n\}} F_1(\hat{\rho})Q_i(\hat{\rho})F_i(\hat{\rho}) = \alpha \tag{A.22b}$$

To find the optimal $\{Q_i\}_{i=2}^n$ that give the maximum $\hat{\rho}=\hat{R}_\alpha$, we define functions $\Phi_i(\rho)\equiv F_1(\rho)F_i(\rho)$ $(i\in\{2,3,\ldots,n\})$. Because Φ_i is continuous, $\lim_{\rho\to-\infty}\Phi_i(\rho)=0$, and $\lim_{\rho\to+\infty}\Phi_i(\rho)=1$, $\exists\hat{\rho}_i\in\mathbb{R}$ s.t. $\Phi_i(\hat{\rho}_i)=\alpha$. Also, we can calculate the index that gives the maximum $\hat{\rho}_i$:

$$i^* \equiv \operatorname*{argmax}_{i\in\{2,3,\ldots,n\}} \hat{\rho}_i \tag{A.23a}$$

$$\hat{\rho}^* \equiv \max_{i\in\{2,3,\ldots,n\}} \hat{\rho}_i \tag{A.23b}$$

We will show that i^* is indeed the optimal decision. Assume some value $\hat{\rho} \in \mathbb{R}$ that satisfies $\hat{\rho} > \hat{\rho}^*$, then Equation (A.22b)'s LHS is larger than α:

$$\begin{aligned}
&\sum_{i\in\{2,3,\ldots,n\}} F_1(\hat{\rho})Q_i(\hat{\rho})F_i(\hat{\rho}) \\
=&\sum_{i\in\{2,3,\ldots,n\}} \Phi_i(\hat{\rho})Q_i(\hat{\rho}) \\
>&\sum_{i\in\{2,3,\ldots,n\}} \Phi_i(\hat{\rho}_i)Q_i(\hat{\rho}) \ (\because \hat{\rho} > \hat{\rho}^* \geq \hat{\rho}_i) \\
=&\,\alpha \sum_{i\in\{2,3,\ldots,n\}} Q_i(\hat{\rho}) \\
=&\,\alpha
\end{aligned} \tag{A.24}$$

Note that i^* is independent from $\{Q_i\}_{i=2}^n$, indicating that after W_1 becomes known, the optimal decision is to select i^* that is defined in Equation (A.23a), which is independent from W_1.

A.4 Analogy to the ellipsoid uncertainty model

$(\hat{\rho}_i, \ldots, \hat{\rho}_i) \in \mathbb{R}^n$ is the solution to the following equations:

$$F_1(W_1)F_i(W_i) = \alpha \tag{A.25a}$$

$$W_1 = W_2 = \cdots = W_n \tag{A.25b}$$

Equation (A.25a) defines a $(n-1)$-dimensional hyperplane in $\mathbb{R}^n$, and Equation (A.25b) defines a one-dimensional line in $\mathbb{R}^n$. Because the optimal reward $\hat{\rho}^*$ is the maximum value of $\hat{\rho}_i$, it is the intersection of the line defined by Equation (A.25b) and the surface of a region surrounded by the hyperplanes of Equation (A.25a):

$$\mathcal{W}_{\text{prospect},1} \equiv \left\{ (W_1, \ldots, W_n) \;\middle|\; \bigwedge_{i\in\{2,3,\ldots,n\}} F_1(W_1)F_i(W_i) \geq \alpha \right\} \tag{A.26}$$

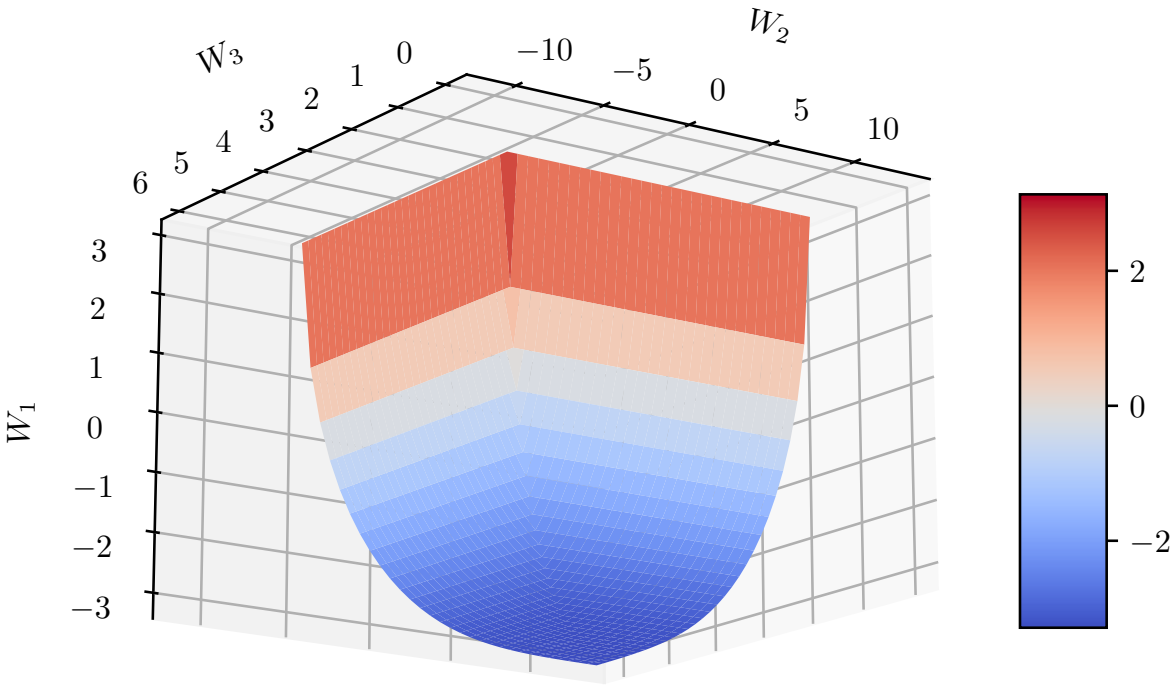


Figure A.1 The region where $F_1(W_1)F_2(W_2) \geq \alpha$ and $F_1(W_1)F_3(W_3) \geq \alpha$. The three random parameters follow normal distributions: $W_1 \sim \mathcal{N}(0, 2^2)$, $W_2 \sim \mathcal{N}(1, 4^2)$, and $W_3 \sim \mathcal{N}(3, 1^2)$. α was set as $\alpha = 0.05$. The color indicates the value of W_1. The intersection point of a line $W_1 = W_2 = W_3$ and the region's surface is $(\hat{\rho}^*, \hat{\rho}^*, \hat{\rho}^*)$.

Figure A.1 shows an example of $\mathcal{W}_{\text{prospect},1}$ in case $n = 3$. It can be seen that the region is asymmetric in that $\mathcal{W}_{\text{prospect},1} \neq \mathcal{W}_{\text{prospect},2} \neq \mathcal{W}_{\text{prospect},1}$ in general, indicating that it is impossible to create an uncertainty region that accurately models the values of all prospecting actions. Therefore, an ellipsoid uncertainty region can be regarded as a symmetric region that can approximately model the values of all prospecting actions.